BEYOND PREDICTION ACCURACY: TARGET-SPACE RECOVERY PROFILES FOR EVALUATING MODEL–BRAIN ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Alignment between vision models and the human brain is often evaluated by how accurately model representations predict brain responses. When two models achieve similar prediction accuracy, this measure does not establish whether they recover the same brain-response structure. We propose a framework for comparing models through their predictive subspaces in the target subject's response space. For each target subject and visual region, we construct a weighted reference from functional magnetic resonance imaging (fMRI) responses across multiple trials of the same images. In the main analysis, this reference averages predictive-subspace projectors obtained by ridge-regularized reduced-rank regression between trial-group means. A recovery profile measures each model's subspace overlap with the weighted reference directions. Differences between profiles obtained from other subjects' responses, matched for prediction accuracy within each target, provide a brain-to-brain benchmark for model differences. We analyzed responses to 515 shared natural images from eight Natural Scenes Dataset (NSD) subjects in early and intermediate visual areas (V1, V2, V3, and hV4). We compared ImageNet-pretrained and randomly initialized models of the same architecture, matched within 0.01 in held-out prediction accuracy. Pretrained models had higher mean recovery-profile scores in 25 of 29 pairs. The mean difference remained positive with an alternative reference based on signal covariance estimated by Generative Modeling of Signal and Noise (GSN). These results show that recovery profiles can reveal differences associated with pretraining, even when prediction accuracy is closely matched.

1 INTRODUCTION

One of the central goals of NeuroAI is to use computational models to understand mechanisms of brain processing and to guide the design of artificial systems (Yamins & DiCarlo, 2016; Zador et al., 2023). For visual processing, this goal is often pursued by comparing how artificial models and the human visual cortex respond to the same images, especially using deep vision models as predictors of visual cortical responses (Yamins et al., 2014; Güçlü & van Gerven, 2015). Encoding analyses evaluate how accurately model representations predict brain responses (Naselaris et al., 2011), whereas representational analyses compare the geometry of model and brain response patterns over a shared stimulus set (Kriegeskorte et al., 2008; Kornblith et al., 2019). Previous studies have linked natural-image training to improved model–brain correspondence (Khaligh-Razavi & Kriegeskorte, 2014; Conwell et al., 2024).

Prediction accuracy does not reveal which response dimensions are recovered. However, it remains unclear: *When pretrained and randomly initialized models achieve similar accuracy for the same target region of interest (ROI), do they recover the same response dimensions?* Answering this question would clarify which aspects of neural representation are captured by learned model features. This requires a method to compare which directions in the target subject's response space are covered by each model's predictions.

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

Brain-to-brain prediction provides an empirical benchmark. We also need to assess these model differences relative to measured brain-to-brain variability, as motivated by the NeuroAI Turing Test (Feather et al., 2025). For a fixed target subject and ROI, other subjects' responses may recover different directions in the target response space. Their coverage differences at similar prediction accuracy can provide an empirical benchmark under the same mapping protocol.

Our approach: compare which response dimensions each prediction recovers. Here, we propose a target-space recovery framework. We estimate reference directions and weights from repeated target responses to held-out images independently of the source being evaluated. For each target subject and ROI, the same reference is used to evaluate all sources predicting that target. We fit model features or other subjects' responses (brain sources) to predict the target on separate training images using the same mapping protocol. Each recovery profile measures a fitted target subspace's reference-weighted coverage of the first k directions as k increases.

Contributions.

1. We propose a target-space recovery profile framework that measures reference-weighted overlap between each source's predictive subspace and directions estimated from repeated responses in the target subject's ROI. We compare differences in mean profile scores between accuracy-matched models with the distribution of within-target differences between accuracy-matched brain sources.
2. We find higher mean recovery-profile scores for ImageNet-pretrained models (Deng et al., 2009; Russakovsky et al., 2015) than for randomly initialized models in 25 of 29 same-architecture pairs whose held-out prediction accuracies differed by at most 0.01. The pretrained-minus-random difference exceeded the pooled 90th percentile of absolute within-target differences between accuracy-matched brain sources in 21 pairs. We also obtain a positive mean pretrained-minus-random difference for the same pairs using signal covariance from Generative Modeling of Signal and Noise (GSN) as the reference (Section 5.2).

2 Related work

Representational dimensions and structured evaluation. Chen & Bonner (2025) identify dimensions shared across visual models and show that a few such model-side dimensions preserve much of their representational similarity to the brain. Canatar et al. (2023) relate neural prediction error to model-dependent stimulus modes of the model-activation Gram matrix. High brain-prediction accuracy can also coexist with model units poorly predicted from neural activity (Muzellec & Kar, 2026). Our reference directions are instead estimated from repeated target responses alone, allowing different sources to be compared in the same target coordinates. Alignment Pattern Analysis (APA) profiles cross-region relationships against a human reference (Höfling et al., 2026). Our profile measures coverage of response directions within a target ROI.

Reliable-signal estimation. Cross-validated principal component analysis (cvPCA) estimates cross-repeat signal power along PCA axes fitted to one trial view (Stringer et al., 2019), while GSN estimates signal and noise covariances (Kay et al., 2025). The reproducible target reference (TargetRef) averages predictive-subspace projectors from regression between trial views. All three use repeat-stable response structure, but TargetRef weights agreement among regularized predictive subspaces rather than signal variance (Appendix C.3).

Brain-to-brain comparisons. Intersubject correlation measures the consistency of stimulus-evoked responses across individuals (Hasson et al., 2004; Nastase et al., 2019). Brain-to-brain mappings can guide the choice of model–brain mappings (Cao & Yamins, 2024). In mouse visual cortex, Nayebi et al. (2023) selected a regression mapping through inter-animal prediction and evaluated model accuracy relative to inter-animal consistency. The inter-animal-transform framework asks whether a model can substitute for a subject under the strictest transform class needed to map accurately between subjects (Thobani et al., 2025). We keep the mapping protocol fixed and compare differences in mean profile scores between accuracy-matched models with those between accuracy-matched brain sources predicting the same target.

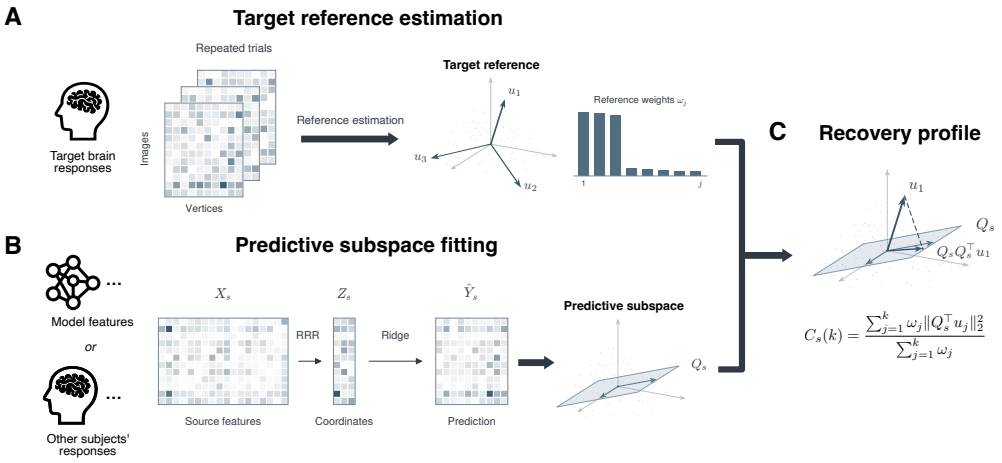

Figure 1: **Overview of the target-space recovery framework.** (A) Repeated target responses to outer-test images define reference directions u_j and weights ω_j. (B) Using outer-training images, ridge-regularized reduced-rank regression (RRR) selects source coordinates Z_s. A ridge map to target responses defines a predictive subspace with orthonormal basis Q_s. (C) $C_s(k)$ measures weighted coverage of the leading k reference directions. For each target unit, all predictive subspaces are evaluated against the same reference in that target's response coordinates.

3 TARGET-SPACE RECOVERY FRAMEWORK

Definitions and overview. For source s, let $X_s \in \mathbb{R}^{n \times p_s}$ contain responses or model features for n images in p_s source coordinates. A *target unit* is a subject–hemisphere–ROI combination, with response matrix $Y \in \mathbb{R}^{n \times q}$ over q retained voxels or vertices. We estimate a reference from the target alone, fit each source separately to predict target responses, and evaluate the induced target subspaces against the same reference (Figure 1).

3.1 TARGET REFERENCE ESTIMATION

We estimate the reference from repeated target responses without using the source being evaluated. It provides orthonormal directions u_j and descending nonnegative weights ω_j with positive total weight. TargetRef and GSN signal covariance both provide such an eigendecomposition, so we compare subspaces and recovery profiles under these two references. We use cvPCA for spectral comparisons because averaging its spectra across trial views does not provide a single target-space eigenbasis (Appendix C.3).

References use held-out target trials in the same target coordinates. Target scaling is fitted to trial-averaged responses for outer-training images. This allows us to change the reference while keeping source fits fixed. Reference hyperparameters are selected within the reference images (Appendix B.2).

TargetRef applies the predictive-subspace representation to repeated-trial responses. For each partition b, we divide each image's trials into two groups and average them to form $Y_b^{(1)}$ and $Y_b^{(2)}$. Ridge-regularized reduced-rank regression (RRR) fits in both directions yield target-side bases $Q_{b,1\to2}$ and $Q_{b,2\to1}$ using the procedure in Section 3.2. We average their projectors over $N_{\text{split}} = 20$ partitions:

$$\text{TargetRef} = \frac{1}{2N_{\text{split}}} \sum_{b=1}^{N_{\text{split}}} \left(Q_{b,1\to2} Q_{b,1\to2}^\top + Q_{b,2\to1} Q_{b,2\to1}^\top \right) = U \,\text{diag}(\omega) U^\top.$$

Each eigenvalue ω_j is the mean squared projection length of direction u_j onto the split-specific predictive subspaces. The leading rank-k reference subspace maximizes mean overlap with these subspaces (Appendix C.2). This construction uses the same predictive-subspace representation for repeat-to-repeat and source-to-target fitting. It is one reference estimator within the framework.

GSN instead weights its directions by estimated signal variance. Regularization and rank selection influence the TargetRef weights.

The twenty partitions are Monte Carlo samples. Ten-partition subsets closely matched the full estimate (Appendix C.4). For either reference, entropy effective rank describes concentration of its own weights:

$$r_{\text{eff}} = \exp\left(-\sum_j p_j \log p_j\right), \qquad p_j = \omega_j / \sum_\ell \omega_\ell.$$

3.2 Predictive subspace fitting

Coverage depends on the span of a fitted mapping, so even weakly predictive directions can increase it. We use RRR to select source coordinates for predicting target responses (Izenman, 1975; Semedo et al., 2019). For each target unit and held-out fold, we construct the predictive subspace in two stages.

First, we select the source dimension k_s and regularization α_{sub} by five-fold inner cross-validation on outer-training images. A one-standard-error rule chooses the smallest rank among settings within one standard error of the best mean validation correlation. The main rank cap is 20. Source-rank controls and target-reference rank-selection sensitivity are reported in Appendices E.1 and C.4, respectively. Training statistics determine feature and response standardization.

Second, we fit a ridge map from the selected source coordinates to target responses, reusing α_{sub}. The range of its transposed coefficient matrix defines an orthonormal target basis Q_s and projector P_s:

$$Q_s = \text{orth}(\hat{B}^\top_{Z\to Y}) \in \mathbb{R}^{q\times k_s}, \qquad P_s = Q_s Q_s^\top.$$

The projector P_s represents the predictive subspace induced by source s in target coordinates, allowing comparison of brain and model sources with different native spaces.

We evaluate prediction accuracy with a separate ridge readout whose regularization α_{read} is selected by inner cross-validation. Accuracy is the mean held-out Pearson correlation across target vertices. This readout does not change Q_s or the recovery profile. The two ridge fits use the same source basis but differ in coordinate standardization and regularization (Appendix B).

Ridge regularization accommodates correlated, high-dimensional features and provides a common linear fitting protocol for brain and model sources (Hoerl & Kennard, 1970). Appendix F.1 examines invariance to orthogonal rotations of standardized source coordinates.

3.3 Recovery profile

For either reference estimator, we project each direction u_j onto source s's induced target subspace. Its squared projection length defines directional coverage:

$$\left\|Q_s^\top u_j\right\|_2^2 = u_j^\top P_s u_j.$$

A value of 1 indicates that u_j lies within the predictive subspace, and 0 indicates orthogonality.

We summarize recovery across leading directions using top-k reference coverage,

$$C_s(k) = \frac{\sum_{j=1}^k \omega_j \|Q_s^\top u_j\|_2^2}{\sum_{j=1}^k \omega_j},$$

the reference-weighted mean of directional coverage over the first k directions. This *recovery profile* can decrease as lower-ranked directions are included. We interpret coverage alongside held-out accuracy because geometric overlap does not measure how accurately responses are predicted.

We summarize the displayed recovery profile by its mean:

$$\text{ProfileMean}_s^{(K)} = \frac{1}{K}\sum_{k=1}^K C_s(k).$$

3.4 Model comparison using brain-to-brain variability

For a fixed target, differences between recovery profiles obtained from other subjects' responses at similar prediction accuracy quantify brain-to-brain variability. We use this variability as an empirical benchmark for model differences.

We compute recovery profiles for each outer fold and then average them. Recovery profiles in the main figures use $K = 10$, which captures most reference weight while keeping the profiles readable (Figure 2B). Sensitivity to K is reported in Appendix Table 8. Normalized ProfileMean and full-spectrum coverage are defined in Appendix B.5.

To assess model differences relative to brain-to-brain variability, we use the pooled 90th percentile of within-target absolute ProfileMean differences between accuracy-matched brain sources. We also compute the percentile using only target blocks containing eligible model pairs, giving each block equal total weight. These are descriptive benchmarks, not significance thresholds or NeuroAI Turing Test pass criteria.

4 Experimental setting

Data and sources. The main analysis uses eight subjects from the Natural Scenes Dataset (NSD) and 515 shared images, with three trials per image selected for reference estimation (Allen et al., 2022). Targets comprise seven visual ROIs in each hemisphere: V1v, V1d, V2v, V2d, V3v, V3d, and hV4. Released GLMsingle single-trial betas provide the responses (Prince et al., 2022). Brain sources are the seven other subjects in the corresponding ROI. Model sources are ImageNet-pretrained and randomly initialized ResNet-18, ResNet-50, VGG-16, and ViT-B/16 (He et al., 2016; Simonyan & Zisserman, 2015; Dosovitskiy et al., 2021). Layer definitions and features are fixed before evaluation (Appendix A.3). The main matched comparisons use random initialization seed 42. Appendix Table 1 summarizes the setting, with seed and dataset controls in Table 9 and Figure 12.

Folds and reference comparisons. We divide the 515 images into five synchronized sets of 103. Each outer fold uses 309 images for source fitting and inner-CV selection of source-subspace and readout hyperparameters, and 103 for testing. The remaining 103 images are excluded from source fitting and selection and used for the disjoint-image reference check (Appendix E.2). Each test-fold reference uses the same 103 images and fixed trial triplets, yielding 560 references across 112 target units. Target scaling is fitted on the 309 training image means and applied to every test trial. We compare the 20-partition reference with official GSN and cvPCA implementations in these coordinates, using GSN covariance without conversion to a correlation matrix (Appendix C.3).

Statistics. We average folds within subject×hemisphere×ROI blocks and describe reference distributions by block medians. The fixed model comparison contains 29 same-architecture pretrained–random pairs in 26 blocks, selected by $|\Delta\text{accuracy}| \leq 0.01$ on five-fold mean accuracy. Pair differences are averaged within blocks, and the 26 block means receive equal weight in 10,000-draw paired-bootstrap 95% confidence intervals. The pooled brain-to-brain benchmark uses 444 accuracy-matched pairs in 110 blocks, weighting pairs equally. Pair identities remain fixed across reference estimators. Appendix Table 8 evaluates accuracy-tolerance sensitivity, and Appendices C.3 and B.7 detail estimator comparisons and benchmark weighting.

5 Results

5.1 A concentrated reference aligned with leading reliable signal

Reference spectra and agreement with GSN. The held-out TargetRef has median effective rank 4.289 across the 112 fold-averaged target blocks. Its first ten directions contain a median 97.3% of reference weight (Figure 2A,B). For comparison, GSN and positive cvPCA spectra have effective ranks 46.065 and 8.654, respectively (Appendix Table 4). These ranks summarize different spectral weights: average projection overlap with predictive subspaces for TargetRef and estimated signal power for the other methods. Despite the broad GSN tail, the leading TargetRef and GSN subspaces agree closely.

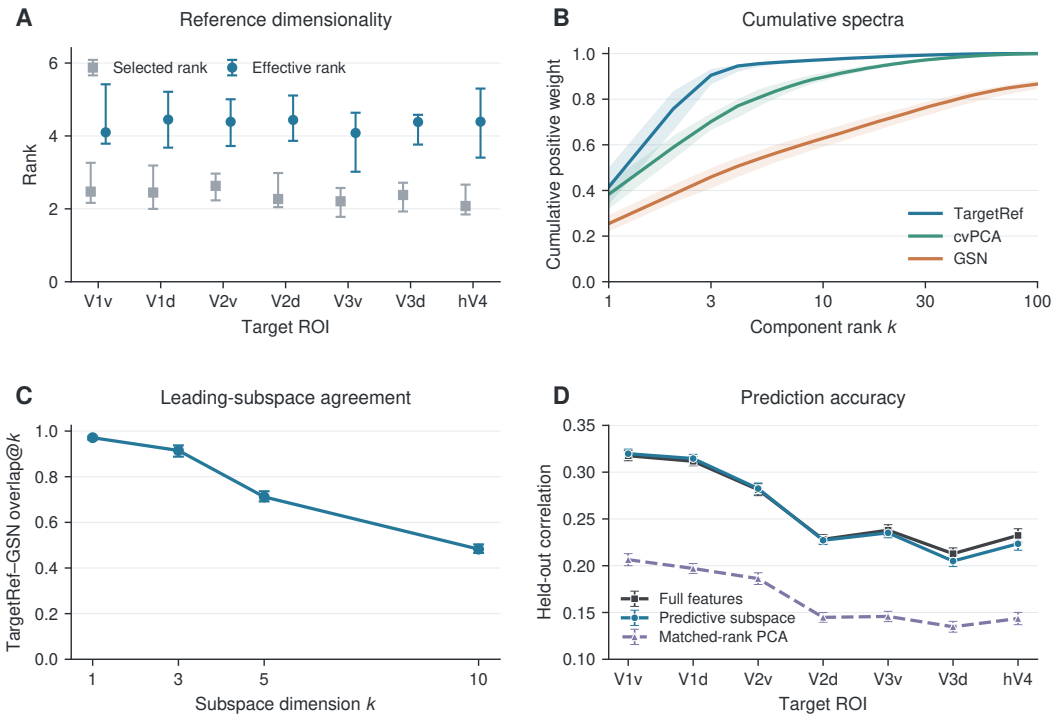

Figure 2: **Reference structure and prediction accuracy.** (A) Selected and effective reference ranks by ROI. Points and bars show medians and interquartile ranges over 16 subject–hemisphere blocks after averaging five folds. Selected rank is first averaged over directional fits. (B) Normalized cumulative spectra, showing medians and interquartile ranges over 112 target blocks. TargetRef weights describe predictive-projector consensus. GSN and positive cvPCA weights describe estimated signal power. (C) TargetRef–GSN subspace overlap, with block medians and 95% bootstrap intervals. Panels A–C use the same 103 held-out images, trial triplets, and target scaling. (D) Prediction accuracy for full features, predictive subspaces selected with rank cap 20, and rank-matched source PCA. All readouts use five-fold inner CV. Points and bars show means and standard errors of the mean (SEMs) over 128 model–target units per ROI after averaging folds. Appendix B.4 gives fitting details.

We compare nonzero positive-semidefinite reference operators A and B in the same target coordinates using Frobenius cosine similarity (hereafter, reference similarity):

$$\mathrm{sim}_F(A, B) = \frac{\mathrm{tr}(AB)}{\sqrt{\mathrm{tr}(A^2)\,\mathrm{tr}(B^2)}}.$$

A value of 1 indicates equality up to positive scale. TargetRef and GSN have reference similarity $0.9082\,[0.9026, 0.9128]$. Their top-1 and top-3 subspace overlaps are 0.9711 and 0.9150, where overlap is the mean squared cosine of the principal angles. The leading three TargetRef directions capture 44.1% of GSN signal variance. Top-10 overlap decreases to 0.4827, indicating greater disagreement among later directions (Figure 2C). Full uncertainty estimates and signal-mass comparisons are in Appendix C.3.

Prediction accuracy after dimension reduction. In a prediction-accuracy control using the same rank cap and inner-CV readout selection, the ROI-mean correlation for selected source coordinates differs from the full-feature correlation by -0.009 to 0.003. Both are higher than source-side matched-rank principal component analysis (PCA) controls (Figure 2D), supporting selection by target prediction rather than source variance.

Repeat count and rank selection. On the same 515 images and fixed target coordinates, pair-specific two-repeat fits have median effective rank 4.567, compared with 5.755 for three-repeat

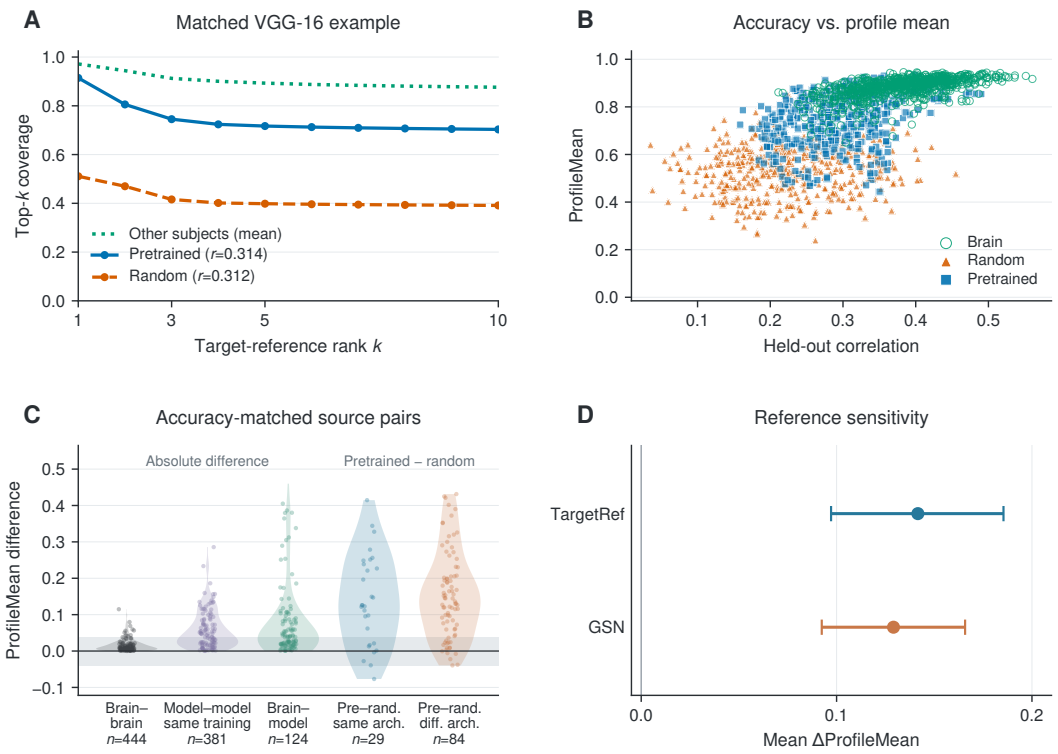

Figure 3: **Models differ in recovered dimensions beyond prediction accuracy.** (A) Profiles for pretrained and randomly initialized VGG-16 in subject 4's right V2v, together with the mean profile from the other seven subjects. (B) Prediction accuracy and ProfileMean for brain and model sources, each averaged over five folds. (C) ProfileMean differences between source pairs with $|\Delta\text{accuracy}| \leq 0.01$. Model pairs are pooled across architectures. The first three categories use absolute differences, and the last two use signed pretrained-minus-random differences. Violins show all pairs, and dots display up to 110 pairs per category. No means are shown in this panel. The gray band spans $\pm$ the pooled 90th percentile of absolute differences between accuracy-matched brain sources. (D) Mean ProfileMean differences for the same 29 same-architecture pretrained–random pairs under TargetRef and GSN, with 95% paired-bootstrap intervals. These estimates weight the 26 eligible target blocks equally. Random models use seed 42.

fits (Appendix Table 6). Both conditions use scaling from the three-repeat image means, so the comparison varies the repeats provided for fitting while keeping target coordinates fixed.

In the matched three-partition control on 103 images, raising the rank cap from 20 to 40 did not change the median effective rank under the one-standard-error rule. At cap 40, selecting the best mean validation score instead increased effective rank, suggesting that one-standard-error selection contributes to reference concentration in this setting. The mean pretrained-minus-random ProfileMean difference for the fixed pairs decreased but remained positive (Appendix C.4).

Recovery from other subjects' responses. Brain-to-brain coverage, averaged across target units and brain sources, declines from 0.951 at $k = 1$ to 0.853 at $k = 10$ as lower-weight reference directions are included. Appendix Figure 9 shows the ROI-resolved profiles.

5.2 Recovery profiles distinguish models with similar prediction accuracy

Accuracy-matched pretrained–random comparisons. In subject 4's right V2v, pretrained and randomly initialized VGG-16 achieved accuracies of 0.3143 and 0.3116, respectively. Their TargetRef ProfileMean scores were 0.7444 and 0.4165, reflecting greater coverage of leading reference directions by the pretrained model (Figure 3A). Across brain and model sources, similar accuracies likewise accompanied different ProfileMean scores (Figure 3B).

Pretrained models had higher ProfileMean in 25 of the 29 fixed accuracy- and architecture-matched pairs. Averaging pair differences within each of the 26 eligible blocks and weighting blocks equally gave a mean pretrained-minus-random difference of $0.1416\,[0.0972, 0.1854]$ (Figure 3D). Within the matching tolerance, pretrained models had higher prediction accuracy in 21 pairs. Among the remaining eight, six still had higher ProfileMean (Appendix Figure 8).

Comparison with brain-to-brain variability. Among brain sources matched for accuracy within each target, the pooled 90th percentile of absolute ProfileMean differences was 0.0387. Twenty-three model pairs exceeded this descriptive threshold in absolute magnitude, and 21 exceeded it in the pretrained direction (Figure 3C). Restricting the benchmark to 115 brain-source pairs in the same 26 eligible target blocks and weighting blocks equally gave a threshold of 0.03355, with the same 23 absolute and 21 positive exceedances (Appendix Table 3).

Model comparisons using GSN. Replacing TargetRef with GSN signal covariance preserved a positive mean pretrained-minus-random ProfileMean difference for the same subspaces and pairs: $0.1291\,[0.0924, 0.1658]$ (Figure 3D). Twenty-six pairs favored pretrained models. In 22 pairs, the difference exceeded the GSN-based 90th percentile of absolute ProfileMean differences between accuracy-matched brain sources (0.0384) in the pretrained direction. Full-spectrum GSN coverage also favored pretrained models, with a difference of $0.0605\,[0.0413, 0.0798]$. The first ten directions accounted for this positive difference when their contribution and the tail contribution were normalized by total GSN signal mass (Appendix C.3).

5.3 Sensitivity and complementary evaluations

Reference images and analysis settings. To test sensitivity to reference-image selection, we rebuilt TargetRef from the reserved 103 images, keeping source fits, target coordinates, and the 29 matched pairs fixed. The mean pretrained-minus-random ProfileMean difference remained positive (0.1454 versus 0.1416), supporting robustness to the choice of reference images (Appendix E.2). The mean difference also remained positive for session/run-balanced partitions, $K = 5, 10, 20$, and accuracy tolerances 0.005–0.03. At the direction level, greater coverage accompanied more accurate held-out prediction (Appendix Table 8). Across all model–target units, the four-seed mean difference was 0.171 (random-seed SD 0.026, Appendix Table 9).

NSD-synthetic. On NSD-synthetic, which contains different stimuli from the same subjects, the corresponding four-seed mean difference was 0.136 across all model–target units without accuracy matching (Appendix Figure 12). The datasets also differ in repeat counts and preprocessing, so this comparison does not isolate the effect of changing stimuli.

Motion sensitivity. To test for motion effects that recur with the same images, we regressed image-locked and within-image motion summaries from trial betas using leave-one-run-out cross-fitting within sessions. We then rebuilt TargetRef with source subspaces and model pairs fixed. The adjusted reference retained 95.4% of baseline weight in its leading ten directions and had mean reference similarity 0.9747 with baseline. The adjustment did not reduce the pretrained-minus-random ProfileMean difference (0.1508 afterward versus 0.1416 beforehand).

We also tested whether reference-direction scores correlated with framewise displacement. To keep direction estimation separate from scoring, we used directions estimated from disjoint image folds. The mean absolute correlation was 0.0582 across the leading three scores. Appendix E.5 reports uncertainty intervals and limitations of the available eye-tracking and physiological data.

Comparison with cross-region alignment patterns. An APA-inspired analysis of the same images and 29 pairs compared model and human cross-region patterns using their correlation, alignment pattern similarity (APS). The pretrained-minus-random APS difference was $-0.059\,[-0.197, 0.089]$, which did not establish a pretrained advantage (Appendix F.2).

First-axis stability and image associations. On images excluded from reference estimation, first-axis response scores correlated across folds at $r = 0.972$ in V1v and 0.996 in the fusiform face area (FFA), averaged within subjects and then across eight subjects. Visual descriptors predicted V1v

scores with test $R^2 = 0.341$. In FFA, facial-landmark annotations added 0.058 in R^2 beyond visual descriptors and person annotations (Appendix Figure 4 and Table 7).

6 DISCUSSION

Higher recovery-profile scores for pretrained models at similar prediction accuracy. ImageNet-pretrained models generally had higher mean recovery-profile scores than randomly initialized models at nearly equal prediction accuracy (Section 5.2). Six of the eight pairs with lower pretrained accuracy also had higher pretrained ProfileMean (Appendix Figure 8). Recovery profiles, therefore, allow us to assess how model representations relate to neural response structure and whether pretraining changes which response dimensions can be predicted from model features.

Model differences relative to brain-to-brain variability. Most matched model pairs exceeded the descriptive brain-to-brain threshold in the pretrained direction (Figure 3C). Matching constrains within-pair accuracy differences, not absolute accuracy, selected rank, or coverage. Because coverage is bounded between 0 and 1, high brain-source coverage may compress the benchmark distribution. Brain-to-brain variability also includes measurement noise, finite-sample effects, and mapping-estimation variability. These comparisons contextualize model differences under the chosen protocol but do not test whether a model can substitute for a biological subject.

Reference estimation. The framework compares source-induced target subspaces against a target-only reference. TargetRef is one way to estimate that reference: it weights agreement among regularized predictive subspaces, whereas GSN weights estimated signal variance (Section 3.1). The positive mean pretrained-minus-random difference under both references supports robustness to this estimator choice (Appendix C.3). TargetRef's greater leading-weight concentration may guide the selection of directions for interpretation, but a smaller effective rank does not establish greater estimation accuracy or stability.

Diagnostic sensitivity across ROIs. We also compared accuracy-matched fully trained architectures in five cortical atlas parcels (PIT, FFC, TE2p, IPS1, and IP1) and functional FFA. These exploratory analyses did not show greater model separation than in early visual ROIs. The atlas parcels had lower TargetRef effective ranks and smaller model differences relative to brain-to-brain variability (Appendix G.5 and Table 10). Both reference-estimation precision and brain-to-brain variability affect this comparison. Additional stimuli and repeats could test whether weakly estimated directions further separate trained models.

Interpretation of leading reference directions. Stable first-axis scores allowed us to compare high-, near-zero-, and low-response images in V1v and FFA (Appendix Figure 5). V1v scores were associated with local brightness variation, while FFA scores were partly explained by facial landmarks on new images (Appendix D). The first reference direction was also largely shared by TargetRef and GSN, with squared axis cosines of 0.970 in V1v and 0.963 in FFA (Appendix Table 7). Recovery profiles quantify how much each model's predictive subspace covers these reference directions, allowing differences associated with pretraining to be compared in the same target space. Specific functional interpretations still require further validation.

Limitations. Recovery measures overlap with the chosen reference under the evaluated ROI, dataset, preprocessing, and mapping class (Sections 3.1–3.2 and 4). Broader brain-likeness requires additional evidence (Feather et al., 2025).

The leading ten directions estimated after motion residualization retained 95.4% of the baseline reference weight (Section 5.3 and Appendix E.5). Trial-aligned eye-tracking and physiological recordings covered too few evaluation trials for comparable residualization analyses. Effects from these sources could still contribute to the reference if they remain in the GLMsingle betas and recur across presentations of the same image.

Our results support evaluating model–brain alignment through prediction accuracy and recovery of reproducible target structure, with accuracy-matched brain-to-brain profiles providing an empirical benchmark for model differences.

AI USE STATEMENT

We used generative AI tools to assist with manuscript writing and polishing and with retrieval and discovery of related work. We also used these tools for code implementation and debugging and for figure and table preparation. The authors carefully reviewed all AI-assisted content and take responsibility for the final manuscript, analyses, and code.

ETHICS STATEMENT

This study analyzes de-identified, previously released human fMRI data and involves no new data collection. The NSD and NSD-synthetic publications describe the original consent and ethics-review procedures (Allen et al., 2022; Gifford et al., 2026). Raw responses, stimulus image files, and third-party model checkpoints are not included in the code distribution. Appendix I lists credits for the stimulus examples reproduced in the paper. The intended use is scientific evaluation of model–brain alignment, not clinical diagnosis or individual identification. Appendix H describes resource terms.

REPRODUCIBILITY STATEMENT

Section 4 specifies the data, folds, and statistical comparisons. Appendices A and B detail preprocessing, feature selection, fitting, metric definitions, and statistical aggregation. Appendix C describes reference construction and estimator comparisons. Appendix E reports robustness analyses. Appendix H describes the accompanying analysis code, required inputs, and resource terms.

REFERENCES

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.

William Brown. Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3):296–322, 1910. doi: 10.1111/j.2044-8295.1910.tb00207.x.

Abdulkadir Canatar, Jenelle Feather, Albert J. Wakhloo, and SueYeon Chung. A spectral theory of neural prediction and alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 47052–47080, 2023. doi: 10.52202/075280-2038.

Rosa Cao and Daniel Yamins. Explanatory models in neuroscience, part 1: Taking mechanistic abstraction seriously. *Cognitive Systems Research*, 87:101244, 2024. doi: 10.1016/j.cogsys.2024.101244.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

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

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*, 15:9383, 2024. doi: 10.1038/s41467-024-53147-y.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

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

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York, 1993.

Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (Python library), 2019. URL `https://github.com/MadryLab/robustness`.

Ky Fan. Maximum properties and inequalities for the eigenvalues of completely continuous operators. *Proceedings of the National Academy of Sciences of the United States of America*, 37(11): 760–766, 1951. doi: 10.1073/pnas.37.11.760.

Jenelle Feather, Meenakshi Khosla, N. Apurva Ratan Murty, and Aran Nayebi. Brain-Model evaluations need the NeuroAI Turing test. *arXiv preprint arXiv:2502.16238*, 2025. doi: 10.48550/arXiv.2502.16238.

Alessandro T. Gifford, Radoslaw M. Cichy, Thomas Naselaris, and Kendrick Kay. A 7T fMRI dataset of synthetic images for out-of-distribution modeling of vision. *Nature Communications*, 17:1589, 2026. doi: 10.1038/s41467-026-69345-9.

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.

Umut Güçlü and Marcel A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience*, 35(27):10005–10014, 2015. doi: 10.1523/JNEUROSCI.5023-14.2015.

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.

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

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

Larissa Höfling, Matthias Tangemann, Lotta Piefke, Susanne Keller, Matthias Bethge, and Katrin Franke. Only brains align with brains: Cross-region alignment patterns expose limits of normative models. In *International Conference on Learning Representations*, 2026.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975. doi: 10.1016/0047-259X(75)90042-1.

Kendrick Kay, Jacob S. Prince, Thomas Gebhart, Greta Tuckute, Jingyang Zhou, Thomas Naselaris, and Heiko H. Schütt. Disentangling signal and noise in neural responses through generative modeling. *PLOS Computational Biology*, 21(7):e1012092, 2025. doi: 10.1371/journal.pcbi.1012092.

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

Nathan C. L. Kong, Eshed Margalit, Justin L. Gardner, and Anthony M. Norcia. Increasing neural network robustness improves match to macaque V1 eigenspectrum, spatial frequency preference and predictivity. *PLOS Computational Biology*, 18(1):e1009739, 2022. doi: 10.1371/journal.pcbi.1009739.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Sabine Muzellec and Kohitij Kar. Reverse predictivity for bidirectional comparison of neural networks and biological brains. *Nature Machine Intelligence*, 8(3):474–488, 2026. doi: 10.1038/s42256-026-01204-0.

Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, 2011. doi: 10.1016/j.neuroimage.2010.07.073.

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.

Aran Nayebi, Nathan C. L. Kong, Chengxu Zhuang, Justin L. Gardner, Anthony M. Norcia, and Daniel L. K. Yamins. Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation. *PLOS Computational Biology*, 19(10):e1011506, 2023. doi: 10.1371/journal.pcbi.1011506.

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: e77599, 2022. doi: 10.7554/eLife.77599.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, pp. 606–610, 2007.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3): 211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Ignacio Santamaría, Louis L. Scharf, Chris Peterson, Michael Kirby, and Joseph M. Francos. An order fitting rule for optimal subspace averaging. In *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 1–4, 2016. doi: 10.1109/SSP.2016.7551843.

João D. Semedo, Amin Zandvakili, Christian K. Machens, Byron M. Yu, and Adam Kohn. Cortical areas interact through a communication subspace. *Neuron*, 102(1):249–259.e4, 2019. doi: 10.1016/j.neuron.2019.01.026.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. doi: 10.48550/arXiv.2508.10104.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Charles Spearman. Correlation calculated from faulty data. *British Journal of Psychology*, 3(3): 271–295, 1910. doi: 10.1111/j.2044-8295.1910.tb00206.x.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D. Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365, 2019. doi: 10.1038/s41586-019-1346-5.

Imran Thobani, Javier Sagastuy-Brena, Aran Nayebi, Jacob S. Prince, Rosa Cao, and Daniel L. K. Yamins. Model-brain comparison using inter-animal transforms. In *Proceedings of Cognitive Computational Neuroscience*, 2025. doi: 10.32470/05ahbfir.

Daniel L. K. Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016. doi: 10.1038/nn.4244.

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.

Anthony Zador, Sean Escola, Blake Richards, Bence Ölveczky, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, James DiCarlo, Surya Ganguli, Jeff Hawkins, Konrad Körding, Alexei Koulakov, Yann LeCun, Timothy Lillicrap, Adam Marblestone, Bruno Olshausen, Alexandre Pouget, Cristina Savin, Terrence Sejnowski, Eero Simoncelli, Sara Solla, David Sussillo, Andreas S. Tolias, and Doris Tsao. Catalyzing next-generation artificial intelligence through NeuroAI. *Nature Communications*, 14:1597, 2023. doi: 10.1038/s41467-023-37180-x.

A DATA AND MODEL REPRESENTATIONS

The primary analysis uses the Natural Scenes Dataset (NSD). We estimate the reproducible target reference (TargetRef) by averaging predictive-subspace projectors over 20 trial partitions in fixed target coordinates. This reference is used in the main text and Appendix unless an alternative is specified. Source subspaces and accuracy-selected model pairs are held fixed when comparing references.

Table 1: **Experimental setting for the primary analysis.**

Component	Setting
Dataset	NSD-core-shared
Subjects	8 target subjects, each with the other 7 subjects as brain sources
Images	515 natural images shared across all subjects
Regions of interest (ROIs)	Visual areas V1–V3, each divided into ventral (v) and dorsal (d) subdivisions, and human visual area 4 (hV4), in both hemispheres
Outer evaluation	5 synchronized image folds: 309 training images, 103 test images, and 103 images for the disjoint-reference check
Target reference	TargetRef from bidirectional ridge-regularized reduced-rank regression (RRR) between trial-group means, using 3 fixed trials per test image and 20 partitions
Model architectures	Residual networks (ResNet-18 and ResNet-50), VGG-16 from the Visual Geometry Group, and a Vision Transformer (ViT-B/16)
Model weights	ImageNet-pretrained and randomly initialized weights for each architecture
Evaluation metrics	Prediction accuracy, recovery profiles, ProfileMean, and ProfileMean normalized by the median across other subjects

A.1 DATASET SELECTION

The primary dataset, NSD-core-shared, is a subset of the Natural Scenes Dataset images that are shared across all eight subjects and have at least three repeated trials for each subject. This rule yields 515 natural images. The robustness dataset, NSD-synthetic, uses the 284 unique synthetic stimuli appearing in the NSD-synthetic experimental ordering for the same eight subjects. In both datasets, image identities are ordered once, and all subjects, hemispheres, ROIs, brain sources, and model sources use that same image order. Five synchronized outer folds are produced from a single fixed random permutation. For NSD-core-shared, each fold has 309 outer-training images and 103 held-out test images. The remaining 103 images, termed the outer-validation set, provide an independent reference for the disjoint-image check (Appendix E.2). They are unused in source fitting and readout selection, including the prediction-accuracy comparison in Figure 2D. For NSD-synthetic, the same five-fold rule yields 170–171 outer-training images and 56–57 held-out test images per fold.

A.2 RESPONSE PREPROCESSING AND ROI EXTRACTION

All functional magnetic resonance imaging (fMRI) responses are released single-trial general linear model (GLM) estimates from NSD and NSD-synthetic, obtained with GLMsingle (Allen et al., 2022; Prince et al., 2022; Gifford et al., 2026). GLMsingle uses GLMdenoise for data-driven nuisance regression and applies ridge regularization. Responses are mapped to the FreeSurfer fsaverage cortical surface template. The main ROIs are V1v, V1d, V2v, V2d, V3v, V3d, and hV4 in each hemisphere. ROI vertices are selected from the corresponding fsaverage ROI labels. The main analyses do not apply an additional reliability threshold or vertex-selection step beyond these ROI masks. Nonfinite response values are set to zero after standardization, and vertices with nonfinite or near-zero standard deviation are assigned a unit standard deviation before z-scoring.

For NSD-core-shared, trial-level beta estimates are loaded session by session. Within each session and hemisphere, each vertex is z-scored across all trials in that session. Trials corresponding to the 515 selected shared images are then retained. Image-level response matrices used for brain and model source-to-target fits are formed by averaging the session-wise z-scored trial responses across all repeats of each selected image. TargetRef fixes three trials per image before partitioning. Trial selection maximizes the number of sessions, then global runs, with a fixed seed breaking ties. For

each held-out image, these session-wise z-scored trials are randomly partitioned into two non-empty groups and averaged to form the two response views.

For NSD-synthetic, trial beta estimates are available as a single synthetic beta series rather than separate NSD-core sessions. Each vertex is therefore z-scored across all synthetic trials for that subject and hemisphere before any repeat averaging. Image-level response matrices are formed by averaging the z-scored trials assigned to each of the 284 synthetic stimuli. These stimuli have two, four, eight, or ten repeats. TargetRef uses all available repeats with the same 20-partition procedure and fixed outer-training target scaling. Because NSD-core-shared and NSD-synthetic differ in stimulus distribution, repeat counts, and response preprocessing, NSD-synthetic is treated as a dataset-shift check rather than a pure replication.

A.3 Model representations

All model representations are fixed before the held-out evaluation. The main analyses use models pretrained on the ImageNet image-classification dataset and architecture-matched randomly initialized controls with initialization seed 42. For each image, selected activations are averaged over spatial dimensions for convolutional layers or over sequence positions for transformer layers, then concatenated across layers. ResNet-18 uses residual stages 1–4 (960 total features), ResNet-50 uses residual stages 1–4 (3840 features), VGG-16 uses five convolutional feature taps spanning early to final convolutional blocks (1472 features), and ViT-B/16 uses transformer blocks 0, 3, 7, and 11 after the model's standard image preprocessing (3072 features). The same fixed feature definitions are used for pretrained and random states. Additional layer and pretraining-objective comparisons are reported in Appendix G.6.

B Predictive fitting and recovery metrics

The equations below refer to a fixed target subject, hemisphere, and ROI. Table 2 summarizes the fitted bases and recovery metrics. Reference estimation is detailed in Appendix C.

Table 2: **Core notation.** Target-space quantities for source s and outer fold f.

Symbol	Meaning
$Q_s^{(f)}$	target-space basis for the source-induced predictive subspace for source s; columns are orthonormal
$P_s^{(f)}$	orthogonal target-space projector for the source-induced predictive subspace $Q_s^{(f)}Q_s^{(f)\top}$
$\mathrm{TargetRef}^{(f)}$	reproducible target reference; an average of orthogonal projection matrices in the target response space
$u_j^{(f)}, \omega_j^{(f)}$	target-reference direction and reference weight/eigenvalue from the eigendecomposition of $\mathrm{TargetRef}$; the directions are columns of an orthogonal eigenvector matrix
$\|Q_s^{(f)\top}u_j^{(f)}\|_2^2$	directional coverage of reference direction j
$C_s^{(f)}(k)$	top-k reference coverage
$\mathrm{ProfileMean}_s^{(K)}$	mean of top-k reference coverage over $k = 1, \ldots, K$
r_{eff}	entropy effective rank of the specified nonnegative spectrum

B.1 Outer folds and standardization

Let the 515 shared images be randomly permuted once using a fixed seed, then divided into five synchronized chunks of 103 images each. For outer fold f, one chunk is used as the held-out test set, the next chunk in cyclic order is reserved for the disjoint-image reference check, and the remaining three chunks form the 309-image training set. The reserved chunk is termed outer validation, but is not used to select source-subspace or accuracy-readout hyperparameters. These are selected via inner cross-validation on the 309 outer-training images. Reference hyperparameters are selected within the images assigned to each reference. The prediction-accuracy comparison in Figure 2D also selects readout regularization by inner cross-validation on the outer-training images and reports

performance on the held-out test chunk. Regression-stage z-scoring statistics are estimated on the relevant training images and applied to validation or held-out images. This is separate from the session-level standardization across all trials described in Appendix A. The target-only 515-image sensitivity analysis uses a fixed scaling from all 515 three-repeat image means and is not used for source evaluation (Section C.4).

B.2 Rank and regularization selection

There are two regularization parameters in the reported analyses. The subspace regularization α_{sub} is used when selecting the reduced-rank predictive subspace and when inducing the source-induced target basis. It is selected from

$$\{10, 10^2, 10^3, 10^4, 10^5, 10^6\}.$$

The readout regularization α_{read} is used only to fit the low-dimensional ridge readout from selected source coordinates to the target responses, for reporting prediction accuracy. It is selected from

$$\{10^{-6}, 10^{-3}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4, 10^5, 10^6\}.$$

Because the selected readout coordinates and target responses are z-scored and the readout rank is small relative to the number of training images, $\alpha_{\mathrm{read}} = 10^{-6}$ is effectively an unregularized least-squares readout in these fits. It is kept as a small positive lower bound for numerical consistency. With the source basis and target projector fixed, we select α_{read} to maximize mean inner-validation correlation. This tunes the accuracy readout without changing the subspaces used for recovery profiles. Changing α_{sub} can change the selected rank and both the source and target subspaces.

For target-reference construction and recovery-profile fits, $(k, \alpha_{\mathrm{sub}})$ is selected by five-fold inner cross-validation over ranks $1 \le k \le k_{\max}$, where $k_{\max} = \min(20, p_s, q, \lfloor n_{\mathrm{train}}/2 \rfloor)$. The reference-sensitivity analyses use the alternative rank grids in Table 5. For source fits, each inner split standardizes source and target responses using inner-training statistics only. Reference fits retain fixed vertex scales and only recenter within each inner split: outer-training scales for held-out references, or all-image scales for the 515-image diagnostics. For each candidate pair, a reduced-rank basis is learned on the inner-training images. Inner-training and inner-validation source responses are projected into the resulting k source coordinates, and a temporary ridge readout from those coordinates to the target responses is fit using the same α_{sub}. The prediction-accuracy comparison in Figure 2D reuses these source bases and selected ranks, with the same cap of 20.

The validation score for a candidate pair is the mean vertex-wise Pearson correlation on the inner-validation images. This temporary readout is used only to choose the predictive subspace. It is not the readout used for the held-out prediction-accuracy metric. A one-standard-error rule is used: find the best mean validation score, retain all $(k, \alpha_{\mathrm{sub}})$ pairs within one standard error of that best score, choose the smallest rank among those retained pairs, and break remaining ties by higher mean validation score and then smaller α_{sub}. The selected source subspace is refit on the full outer-training set. Reference fits use all images assigned to that reference. Readout regularization is selected separately via five-fold inner cross-validation, choosing the α with the highest mean validation correlation and breaking ties in favor of the smaller α.

B.3 Source-induced target subspaces

For source s (model features or another subject's responses), training data X_s and Y define the ridge-regularized reduced-rank regression (RRR) problem

$$\min_{B:\mathrm{rank}(B)\le k} \|Y_z - X_{s,z}B\|_F^2 + \alpha_{\mathrm{sub}}\|B\|_F^2,$$

where $X_{s,z}$ and Y_z are z-scored using training-only statistics, and $\|\cdot\|_F$ denotes the Frobenius norm. After selecting $(k, \alpha_{\mathrm{sub}})$, the fit is recomputed on the full outer-training set. In standardized coordinates, let

$$C = X_{s,z}^\top Y_z, \qquad R = X_{s,z}^\top X_{s,z} + \alpha_{\mathrm{sub}} I, \qquad W_\alpha = R^{-1}C.$$

We first compute the leading eigenvectors of the target-space matrix $C^\top W_\alpha$. If $C^\top W_\alpha = V\Lambda V^\top$, multiplication by W_α gives the unnormalized standardized source directions:

$$A_{\mathrm{sub}} = W_\alpha V_{1:k}.$$

We obtain an orthonormal basis for their column space using QR decomposition, $Q_{s,f}^{X,z} = \mathrm{orth}(A_{\mathrm{sub}})$. Dual or smaller eigensystems are used when the feature dimension exceeds the sample size. The persisted source basis is

$$Q_{s,f}^{X} = \mathrm{orth}\Big(D_X^{-1} Q_{s,f}^{X,z}\Big),$$

where D_X contains the training feature standard deviations. To construct the saved source-induced target projectors, we use

$$Z_s^{\mathrm{sub},(f)} = X_{s,z}^{(f)} Q_{s,f}^{X}.$$

The accuracy branch below uses raw-source projections followed by coordinate standardization. Both conventions are held fixed across the reference-only comparisons. The source-induced target basis is the range of the supervised readout from selected source coordinates to target responses. We estimate this readout with ridge regression from Z^{sub} to the z-scored target responses using the same selected α_{sub}, then take the orthonormal column space of the transposed coefficient matrix:

$$Q_s^{(f)} = \mathrm{orth}\Big(\hat{B}_{Z\to Y}^{(f)\top}\Big).$$

The columns of $Q_s^{(f)}$ are orthonormal, and $P_s^{(f)} = Q_s^{(f)} Q_s^{(f)\top}$ is the target-space projector. Target-basis induction reuses the selected α_{sub}. Prediction accuracy uses the separately selected α_{read}.

B.4 Prediction accuracy

After $Q_s^{(f)}$ is defined, prediction accuracy is measured by a separate ridge readout from the selected source coordinates to the full target response. The readout uses coordinates

$$Z_s^{\mathrm{read},(f)} = X_s^{(f)} Q_{s,f}^{X},$$

with coordinate and response z-scoring re-estimated inside each readout cross-validation fold. Let $Z_{\mathrm{read},z}^{(f)}$ denote these selected coordinates after the relevant training-fold z-scoring. The fitted readout is

$$\hat{C}_{\mathrm{read}}^{(f)} = \arg\min_{C} \|Y_z^{(f)} - Z_{\mathrm{read},z}^{(f)} C\|_F^2 + \alpha_{\mathrm{read}} \|C\|_F^2.$$

The readout parameter α_{read} is selected by five-fold inner cross-validation on the outer-training images using the grid above, and the selected readout is refit on all 309 outer-training images. Figure 2D uses the same inner folds and grid for full features, the saved predictive source coordinates, and rank-matched source PCA. PCA uses the leading variance-ordered components of mean-centered outer-training features, with rank matched to the corresponding predictive subspace. Both low-dimensional bases are fixed during readout selection, and coordinate z-scoring is re-estimated within each inner fold. The 103 outer-test images are used only for evaluation. Prediction accuracy is the mean Pearson correlation across target vertices between the z-scored outer-test responses and the z-scored predictions. It is not computed by reading out only the source-induced target basis, which would evaluate a projected response rather than the full target response. This readout affects prediction accuracy but does not change $Q_s^{(f)}$, the recovery profile, or its summaries. Thus, prediction accuracy evaluates the prediction of the full target response, whereas the recovery profile evaluates which target-reference dimensions are spanned by the source-induced predictive subspace.

B.5 Recovery profiles and scalar summaries

Let $u_j^{(f)}$ and $\omega_j^{(f)}$ be the eigenvectors and nonnegative eigenvalues of the reference in fold f. For a brain or model source s, directional coverage is

$$\Big\| Q_s^{(f)\top} u_j^{(f)} \Big\|_2^2.$$

This squared projection length lies between 0 and 1. A value of 1 indicates that the reference direction lies within the source-induced target subspace, and 0 indicates orthogonality. The main figures use the top-k reference coverage profile

$$C_s^{(f)}(k) = \frac{\sum_{j=1}^{k} \omega_j^{(f)} \|Q_s^{(f)\top} u_j^{(f)}\|_2^2}{\sum_{j=1}^{k} \omega_j^{(f)}}.$$

For a fixed reference, top-k reference coverage is unchanged by orthogonal changes of basis within the source-induced target subspace. Because it is a weighted average of directional coverage with nonnegative reference weights, it also lies between 0 and 1. A value near 1 indicates strong overlap with the reference-weighted leading k directions, whereas a value near 0 indicates little overlap. Individual axes can rotate within nearly tied eigenspaces, and top-k subspace stability also depends on the eigenvalue gap at the kth boundary. The curve need not increase with k: adding a lower-ranked direction can lower the weighted average if that direction is weakly covered. Curves are averaged by within-reference rank, so "rank 1" means the leading direction in each fold-specific target reference, not the same vector across subjects or folds.

These are coverage metrics, not direction-wise held-out prediction metrics. For example, directional coverage equals 1 when the induced target subspace contains $u_j^{(f)}$, even if the scalar coordinate $Yu_j^{(f)}$ is predicted weakly on held-out images. A map spanning all q target dimensions has projector I_q and coverage of 1 for every direction, regardless of prediction accuracy. Conversely, prediction accuracy can be high because of response variance outside the leading displayed reference prefix. We therefore interpret recovery profiles together with the fitted rank and prediction accuracy.

The full recovery profile is defined for all q reference directions of $\mathrm{TargetRef}^{(f)}$, where q is the number of retained target vertices for the target subject, hemisphere, and ROI. Thus, the maximum profile length is not fixed globally: in the NSD-core-shared analyses, it ranges from 307 to 1117 directions across target units, with a median of 604.5. For any fixed target fold, however, the reference directions and q are identical for all brain and model sources, so their profiles are directly comparable. The source-induced target basis $Q_s^{(f)}$ can have a much smaller cross-validated rank than q. Recovery at later reference directions is then the squared projection of those directions onto that lower-dimensional source-induced basis.

Profile summaries (ProfileMean). Top-k reference coverage is first averaged across outer folds:

$$C_s(k) = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} C_s^{(f)}(k).$$

The ProfileMean summary used in the main figures is

$$\mathrm{ProfileMean}_s^{(K)} = \frac{1}{K} \sum_{k=1}^{K} C_s(k).$$

Since it is an average of top-k reference coverage, $\mathrm{ProfileMean}$ also lies between 0 and 1. It summarizes the overall height of the displayed profile, not the curve's shape. All main recovery-profile figures use $K = 10$, i.e., the first ten eigendirections of each held-out target reference. This is a reporting convention chosen after inspecting the reference spectra, not a hard rank imposed on the subspace fits. Table 8 reports sensitivity to this choice. The full $\mathrm{TargetRef}$ is computed over all retained target vertices, and $(k, \alpha_{\mathrm{sub}})$ selection is allowed to choose ranks up to $k_{\max}$ as described above.

For target subject t, normalized ProfileMean is a source's ProfileMean divided by the median ProfileMean obtained from the other seven subjects:

$$\frac{\mathrm{ProfileMean}_s^{(K)}}{\mathrm{med}_{d \neq t}\, \mathrm{ProfileMean}_d^{(K)}},$$

where $d \neq t$ indexes the other subjects used to predict the target responses in the corresponding hemisphere and ROI. A value of 1 indicates that a source matches the median ProfileMean from the other subjects for that target subject, hemisphere, and ROI. Values above 1 indicate greater ProfileMean than this median. They do not imply that the full source representation is more brain-like than a human brain. The denominator summarizes recovery from other subjects for the same target, not total response reliability. We also compute full-spectrum reference coverage as

$$\frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \frac{\mathrm{tr}\Big(P_s^{(f)}\, \mathrm{TargetRef}^{(f)}\Big)}{\mathrm{tr}\Big(\mathrm{TargetRef}^{(f)}\Big)}.$$

Within each fold, full-spectrum coverage averages all q directional-coverage values using the reference eigenvalues as weights. It is the quantity plotted for principal component analysis (PCA) and null controls in Figure 6F.

B.6 AGGREGATION AND UNCERTAINTY

Unless otherwise specified, confidence intervals for model comparisons use a target-unit block bootstrap (Efron & Tibshirani, 1993). A target unit is a subject–hemisphere–ROI combination. When a target unit is sampled, all model entries, architectures, and states associated with that target unit are resampled together. This preserves the dependence among model comparisons that share the same target responses. The paired higher-minus-early comparison in Appendix G.5 instead averages ROIs within each subject–hemisphere unit before bootstrapping those 16 units.

We first average folds within each target unit. Reference-only summaries report medians across target units unless stated otherwise. For the fixed accuracy-matched model comparison, eligible pair differences are averaged within each target unit, and the resulting 26 block means receive equal weight. The same block indices are used for paired conditions. Confidence intervals use 10,000 percentile-bootstrap draws. Seeds, repeat subsets, and folds are not counted as independent samples.

B.7 BRAIN-TO-BRAIN BENCHMARK SENSITIVITY

The pooled benchmark uses 444 accuracy-matched brain-source pairs in 110 target blocks. To match the target distribution of the 29 model pairs, we also restrict these brain pairs to the 26 eligible model-comparison blocks. All 26 contain eligible brain pairs, giving 115 pairs. Each pair in block b receives weight $1/n_b$, where n_b is the number of eligible brain pairs in that block. The block-weighted 90th percentile is the smallest absolute difference at which the cumulative normalized weight reaches 0.9. The pooled percentile uses linear interpolation between ordered pair differences. Pair identities are unchanged across reference estimators.

Table 3 reports both benchmarks. The restricted analysis addresses target-block composition and unequal pair counts. It does not match absolute accuracy, selected rank, or coverage levels between brain and model sources. Threshold exceedances remain descriptive.

Table 3: **Brain-to-brain benchmarks for the same 29 model pairs.** Counts indicate absolute or signed pretrained-minus-random ProfileMean differences above the threshold. Pooled benchmarks use 444 brain pairs in 110 blocks. Restricted benchmarks use 115 pairs in the 26 model-comparison blocks, with equal total weight per block.

Reference	Brain-pair weighting	90th percentile	Absolute	Pretrained direction
TargetRef	Pooled pairs	0.03872	23/29	21/29
TargetRef	Eligible blocks	0.03355	23/29	21/29
GSN	Pooled pairs	0.03836	23/29	22/29
GSN	Eligible blocks	0.03131	25/29	24/29

C TARGET-REFERENCE ESTIMATION AND VALIDATION

C.1 TARGETREF CONSTRUCTION

For each trial partition b, responses to an image are divided into two non-empty groups and averaged. We sample $N_{\text{split}} = 20$ partitions using only outer-test images. TargetRef applies the outer-training target scaling to all views and retains that scaling throughout projector construction. Partition-count sensitivity is reported in Section C.4.

The two within-target fits induce target bases $Q^{(f)}_{b,1\to2}$ and $Q^{(f)}_{b,2\to1}$ using the same (k, α_{sub}) selection rule in Appendix B, with the images available to that within-target fit serving as the data pool for

inner cross-validation. Their projectors are averaged:

$$\mathrm{TargetRef}^{(f)} = \frac{1}{2N_{\mathrm{split}}} \sum_{b=1}^{N_{\mathrm{split}}} \left(Q_{b,1\to2}^{(f)} Q_{b,1\to2}^{(f)\top} + Q_{b,2\to1}^{(f)} Q_{b,2\to1}^{(f)\top} \right).$$

This is the reproducible target reference. Its eigendecomposition is

$$\mathrm{TargetRef}^{(f)} = U^{(f)} \operatorname{diag}\left(\omega^{(f)}\right) U^{(f)\top}.$$

Here, $U^{(f)}$ is an orthogonal eigenvector matrix, $U^{(f)\top}U^{(f)} = I$, because $\mathrm{TargetRef}^{(f)}$ is symmetric. The reference direction $u_j^{(f)}$ is column j of $U^{(f)}$, and $\omega_j^{(f)}$ is the corresponding eigenvalue. The eigenvalues are nonnegative because TargetRef is an average of projection matrices. If only the leading reference directions are stored or plotted, they are the corresponding orthonormal columns of this full eigenvector matrix. For spectra and weighted profiles, numerical negative eigenvalues are set to zero, and reference weights are normalized as $p_j = \omega_j / \sum_\ell \omega_\ell$. The prefix ratios are unchanged if raw nonnegative ω values are used.

The rank-matched PCA and random-subspace controls use the rounded mean selected rank within each fold,

$$k_{\mathrm{ref}}^{(f)} = \mathrm{round}\left(\frac{1}{2N_{\mathrm{split}}} \sum_{b=1}^{N_{\mathrm{split}}} \left(k_{b,1\to2}^{(f)} + k_{b,2\to1}^{(f)} \right) \right),$$

clipped to be at least 1. Here, $k_{b,1\to2}^{(f)}$ and $k_{b,2\to1}^{(f)}$ are the ranks selected for the two prediction directions. This rank sets the dimension of the control subspaces, without truncating TargetRef.

Effective rank. For nonnegative weights ω_j with positive total weight, entropy effective rank (Roy & Vetterli, 2007) is

$$r_{\mathrm{eff}} = \exp\left(-\sum_j p_j \log p_j \right), \qquad p_j = \omega_j / \sum_\ell \omega_\ell.$$

Effective rank equals r when r directions have equal positive weights and all remaining weights are zero. We apply this definition to each reference's nonnegative eigenvalues. For cvPCA, only the positive part of the spectrum enters this summary.

For TargetRef, effective rank is first computed within each fold, then averaged over five folds. Its median across 112 target blocks is 4.289, and its median top-ten weight is 97.3%.

C.2 Properties of TargetRef

Repeat-stable covariance. Suppose two trial-averaged views of the same images satisfy

$$Y^{(1)} = S + E^{(1)}, \qquad Y^{(2)} = S + E^{(2)},$$

where S is the image-locked repeat-stable response. If the zero-mean noise terms are independent of S and each other, with finite second moments, then

$$\mathrm{Cov}\left(Y^{(1)}, Y^{(2)}\right) = \mathrm{Cov}(S).$$

The covariance expansion gives this identity because all terms involving noise vanish. It expresses the repeat-stable information underlying split-half comparisons (Spearman, 1910; Brown, 1910). Shared session effects and stimulus-locked nuisance can violate the independence assumptions or contribute to S. The finite-sample predictive subspaces also depend on regularization, rank selection, and estimation error (Section C.3).

Projector averaging. For a fixed target fold, let $P_1, \ldots, P_M$ be the $M = 2N_{\mathrm{split}}$ projectors from the bidirectional trial-group fits, with $\bar{P} = \mathrm{TargetRef}^{(f)} = M^{-1} \sum_{\ell=1}^{M} P_\ell$. For any rank-$k$ orthogonal projector P,

$$\frac{1}{M} \sum_{\ell=1}^{M} \mathrm{tr}(P P_\ell) = \mathrm{tr}(P \bar{P}).$$

Ky Fan's maximum principle (Fan, 1951) gives a maximizer spanning the leading k eigenvectors of $\bar{P}$. This is the projection-matrix averaging used in consensus subspace estimation (Santamaría et al., 2016). The maximizing subspace need not be unique when the eigenvalues at the rank boundary are tied. Each eigenvalue $\omega_j = u_j^\top \bar{P} u_j$ measures the mean squared projection of its direction onto the fitted predictive subspaces. Recovery profiles use the full weighted reference.

Sampling trial partitions. Conditional on the observed responses, independently sampled partitions estimate the mean predictive projector under the chosen partition distribution. Averaging more partitions reduces Monte Carlo variance without adding stimulus repeats. This averaging does not establish consistency for a population signal subspace. Section C.4 reports empirical sensitivity to partition count and separately examines repeat count.

C.3 Comparison with GSN and cvPCA

Matched data and coordinates. We compared TargetRef with Generative Modeling of Signal and Noise (GSN) (Kay et al., 2025) and cross-validated principal component analysis (cvPCA) (Stringer et al., 2019) using the same response tensor of 103 held-out images $\times$ three trials $\times$ q target vertices in each outer fold. Vertex means and standard deviations were fitted to the 309 outer-training image means and applied to every held-out trial. We constructed TargetRef by averaging predictive projectors across 20 trial partitions in these fixed target coordinates. Saved source-induced target subspaces were held fixed. GSN estimated signal and noise covariances with shrinkage. We used its final signal covariance directly, without conversion to a correlation matrix. For cvPCA, we averaged six directional spectra: both directions of each singleton-versus-mean-of-the-other-two partition. Both used the official implementations listed in Table 12.

Relation to the signal covariance. Under the independent additive-noise model in Section C.2, all three methods use responses that share an image-locked component S. Let Σ_S denote its covariance and assume a common single-trial noise covariance Σ_E. For disjoint trial views and an average $\bar{Y}_m$ over m trials,

$$\mathrm{Cov}(Y^{(1)}, Y^{(2)}) = \Sigma_S, \qquad \mathrm{Cov}(\bar{Y}_m) = \Sigma_S + \Sigma_E/m.$$

These identities explain their common dependence on repeat-stable responses, but the estimated quantities differ.

GSN: signal covariance. GSN uses the identity $\Sigma_S = \mathrm{Cov}(\bar{Y}_m) - \Sigma_E/m$. It estimates noise from within-image trial variation and data covariance from trial-averaged responses, then uses shrinkage and positive-semidefinite constraints to obtain its final covariance estimates. Its signal covariance has the decomposition

$$G = V \,\mathrm{diag}(g) V^\top \in \mathbb{R}^{q \times q}, \qquad g_j \geq 0.$$

The eigenvectors v_j and eigenvalues g_j provide target-space directions and estimated signal-variance weights for the recovery profile.

cvPCA: cross-validated spectrum. Let $A, B \in \mathbb{R}^{n \times q}$ be two trial views centered across the same n images, and let $v_j^{(A)}$ be the jth PCA axis of A. The implementation computes

$$\hat{\lambda}_j^{A \to B} = (A v_j^{(A)})^\top (B v_j^{(A)}).$$

Division by $n - 1$ gives cross-validated covariance along that axis. This common factor cancels in normalized spectra and effective rank. Thus, cvPCA does use axes, but they are fitted to one noisy view rather than to an estimated signal covariance. We average these spectral values by component rank across six directed views, whose PCA bases can differ. This averaged spectrum does not specify a single target-space eigenbasis or a positive-semidefinite covariance operator. We therefore use it for spectral comparison, and use GSN for comparisons of reference directions and recovery profiles. Constructing a covariance from cvPCA outputs would require additional choices about axes, negative values, and aggregation.

Table 4: Spectral concentration on matched held-out responses. Effective ranks are block medians with bootstrap intervals. The final column is the median fraction of each estimator's own positive weight in its first ten components, with normalization performed before fold averaging.

Reference	Entropy effective rank	Weight in first ten
TargetRef	$4.289\ [4.111, 4.504]$	0.9729
cvPCA, positive spectrum	$8.654\ [8.197, 9.166]$	0.8962
GSN signal covariance	$46.065\ [42.449, 47.638]$	0.6285

TargetRef: predictive-subspace consensus. For a population reduced-rank ridge step from one trial view to another, the target-side matrix used for rank reduction is

$$\Sigma_{21}(\Sigma_{11} + \lambda I)^{-1}\Sigma_{12} = \Sigma_S(\Sigma_S + \Sigma_E/m + \lambda I)^{-1}\Sigma_S,$$

where $\Sigma_{ab} = \mathrm{Cov}(Y^{(a)}, Y^{(b)})$, m is the input-view trial count, and λ is the covariance-scale ridge penalty. This expression describes the initial rank-reduction step, not the final reference. Our two-stage fitting procedure induces a target projector for each direction of each partition (Appendix B). TargetRef averages these projectors and uses their eigendecomposition to provide reference directions and weights. Each weight $\omega_j = u_j^\top \mathrm{TargetRef}\, u_j$ is the mean squared projection of u_j onto the fitted subspaces, not an estimate of variance along u_j. Within-view noise, regularization, rank selection, and partition agreement therefore affect TargetRef even when the cross-view covariance equals Σ_S.

Spectral concentration and geometry. Reference-only statistics in this section were averaged over five folds within each subject $\times$ hemisphere $\times$ ROI block, followed by the median across 112 blocks. Intervals below are 95% percentile intervals from $10{,}000$ block-bootstrap draws. Paired comparisons reuse the same bootstrap block indices. Table 4 compares concentration under the three estimators. cvPCA's negative entries were retained for diagnostics and comprised 0.144% of total absolute spectral mass. Only positive entries enter its effective rank and cumulative-weight summaries. The spectral weights have different meanings: average projection overlap with predictive subspaces for TargetRef, signal variance for GSN, and cross-validated signal power for cvPCA. Accordingly, these effective ranks describe estimator-specific concentration rather than a common intrinsic cortical dimensionality.

For nonzero positive-semidefinite operators in the same coordinates, we use Frobenius cosine similarity (hereafter, reference similarity):

$$\mathrm{sim}_F(A, B) = \frac{\mathrm{tr}(AB)}{\sqrt{\mathrm{tr}(A^2)\,\mathrm{tr}(B^2)}}. \tag{1}$$

A value of 1 indicates equality up to positive scale. Top-k subspace overlap is

$$\frac{\|U_{A,1:k}^\top U_{B,1:k}\|_F^2}{k},$$

where $U_{A,1:k}$ and $U_{B,1:k}$ contain the leading orthonormal eigenvectors of each operator. This overlap is the mean squared cosine of the principal angles between the two subspaces. TargetRef–GSN reference similarity was $0.9082\ [0.9026, 0.9128]$. Overlap at ranks 1, 3, and 10 was respectively $0.9711\ [0.9628, 0.9763]$, $0.9150\ [0.8880, 0.9372]$, and $0.4827\ [0.4660, 0.5035]$. The first three and ten TargetRef directions captured $44.1\%\ [43.1\%, 45.2\%]$ and $52.7\%\ [51.3\%, 54.3\%]$ of total GSN signal variance. Thus, the leading subspaces are closely aligned, while later directions and spectral weights differ.

Fixed accuracy-matched model comparison. The model comparison used 29 same-architecture pretrained–random pairs in 26 eligible blocks, selected only by five-fold mean prediction accuracy ($|\Delta\mathrm{accuracy}| \le 0.01$; random seed 42). Pair identities were unchanged across reference estimators. Differences were first averaged over folds and then over eligible architectures within each block. The 26 blocks received equal weight in inference. The mean pretrained-minus-random ProfileMean difference was $0.1416\ [0.0972, 0.1854]$ under TargetRef and $0.1291\ [0.0924, 0.1658]$ under GSN. These differences describe the accuracy-matched subset rather than all target blocks.

To assess the broader GSN spectrum, let $g_j \geq 0$ and v_j be its eigenvalues and eigenvectors, and let $d_j = \|Q^\top v_j\|_2^2$ for a saved source-induced target basis Q. With one common denominator,

$$C_{\leq 10}(Q) = \frac{\sum_{j \leq 10} g_j d_j}{\sum_j g_j}, \qquad C_{>10}(Q) = \frac{\sum_{j>10} g_j d_j}{\sum_j g_j}, \qquad C_{\text{full}} = C_{\leq 10} + C_{>10}. \tag{2}$$

Each contribution was computed within a fold before averaging. The corresponding pretrained-minus-random differences were 0.0612 $[0.0421, 0.0805]$ for $C_{\leq 10}$ and -0.0008 $[-0.0009, -0.0007]$ for $C_{>10}$, yielding a full-spectrum coverage difference of 0.0605 $[0.0413, 0.0798]$. Point estimates and paired bootstrap draws are additive. Marginal interval endpoints need not be. The positive full-spectrum difference is concentrated in the first ten GSN components. ProfileMean uses separately normalized prefixes, so its magnitude is not directly comparable to full coverage.

C.4 Sensitivity to repeats, images, and rank selection

Table 5 distinguishes the fixed-coordinate reference analyses, including the matched three-partition RRR reference (RRR3). All RRR reference fits used $\alpha_{\text{sub}} \in \{10, 10^2, \ldots, 10^6\}$ and five-fold inner cross-validation. The saved source fits and accuracy-selected pairs were unchanged.

Table 5: Reference-analysis conditions. The 515-image analyses are target-only diagnostics. The one-standard-error rule selects rank unless stated otherwise. $K_{40} = \{1, 2, 3, 5, 8, 10, 15, 20, 30, 40\}$ is expanded with $60, 80$ only if the initial selected rank reaches 40.

Reference	Images and target scaling	Projector aggregation	Candidate ranks
TargetRef	103 held-out images; scaling from 309 training image means	Both directions of 20 random trial partitions	$1, \ldots, 20$
RRR3	Same as TargetRef	Both directions of three singleton-versus-pair partitions	$1, \ldots, 20$; controls to 40
515-image R_2	All 515 images; scaling from their three-repeat means	Two directions per trial pair; scalar metrics averaged across pairs	K_{40}, with conditional expansion
515-image R_3	Same coordinates as R_2	Six singleton-versus-pair projectors; also partition-first diagnostics	Same as R_2

Repeat sensitivity in fixed coordinates. Using all 515 images in fixed target coordinates, we constructed a separate two-repeat reference for each pair $(a, b) \in \{(1, 2), (1, 3), (2, 3)\}$:

$$R_2^{(ab)} = \tfrac{1}{2}(P_{a \to b} + P_{b \to a}). \tag{3}$$

The omitted repeat was excluded from that pair's predictive fits. The common target scaling was fitted to all 515 three-repeat image means, so this comparison varies repeats provided for fitting conditional on fixed coordinates, not repeat availability throughout preprocessing. The three-repeat reference R_3 averaged six directed projectors from the three singleton-versus-mean-of-two partitions. Effective rank and reference similarity were computed separately for each two-repeat pair, then averaged within each target block. The overlapping repeat subsets were not independent bootstrap observations. Table 6 reports block medians and paired differences of medians over 112 blocks. Median pair-averaged R_2–consensus R_3 reference similarity was 0.9249 $[0.9154, 0.9327]$.

The three-repeat reference retained a higher effective rank when projector counts were matched. The comparison covers only two versus three repeats and does not establish a scaling law or saturation. Because the 515-image references include outer-training images, they are used only for target-reference diagnostics. No model profiles are evaluated against them.

The 103-image RRR3 effective rank was 3.691 $[3.352, 3.875]$, compared with 5.755 $[5.591, 5.877]$ for 515 images. Target scaling and candidate-rank grids also differ between these analyses (Table 5), so this comparison does not isolate image count alone.

Table 6: Two-versus-three-repeat reference fits on the same 515 images in fixed target coordinates. The partition-first diagnostic scores each two-direction partition separately before averaging its scalar statistics, matching the number of projectors contributing to each comparison.

Reference or paired difference	Effective rank or rank difference
Two-repeat R_2, pair-first	$4.5666\ [4.0056, 4.8685]$
Three-repeat R_3, consensus	$5.7552\ [5.5909, 5.8771]$
Consensus R_3 minus pair-first R_2	$1.1886\ [0.9545, 1.6489]$
Three-repeat R_3, partition-first	$5.3662\ [4.7310, 5.5239]$
Partition-first R_3 minus pair-first R_2	$0.7997\ [0.5754, 1.1083]$

Rank-selection control. A separate control used the *matched three-partition RRR3 reference* on 103 held-out images. RRR3 had reference similarity 0.9789 with TargetRef, and their fold-averaged ProfileMean values across all source–target units correlated at $r = 0.9948$. Raising the candidate-rank cap from 20 to 40 while retaining the one-standard-error rule left its median effective rank at $3.691\ [3.352, 3.875]$, although individual fits could differ. The paired difference of medians was $0.000\ [0.000, 0.031]$. Selecting the highest mean validation correlation on the cap-40 grid instead increased effective rank to $12.876\ [11.627, 14.357]$. The fixed-pair pretrained-minus-random ProfileMean difference remained positive at $0.0915\ [0.0582, 0.1250]$, using the same equally weighted mean across 26 target blocks. These controls distinguish the effects of the candidate-rank cap and rank-selection policy.

Number of sampled partitions. Holding the response data and saved rank choices fixed, we formed references from subsets of the twenty trial partitions used to construct TargetRef, keeping both directions of each partition together. Scalar statistics were computed for each subset before averaging within folds and target blocks. The median effective rank increased from 3.082 for one partition to 4.289 for twenty. Median reference similarity to the full 20-partition TargetRef reached 0.9918 at five partitions and 0.9972 at ten. These values describe convergence within the saved pool of partitions, separate from the effect of acquiring additional stimulus repeats.

D Stability and interpretation of leading reference directions

D.1 Frozen directions and fold stability

We examined the first TargetRef direction in ventral V1 (V1v) and the independently localized fusiform face area (FFA) in both hemispheres of all eight Natural Scenes Dataset (NSD) subjects. Each saved direction was estimated from its fold's 103 reference images, using target scaling fitted on 309 training images. We projected responses to nonshared NSD images onto these frozen directions without refitting the reference. Eligible images had at least three trials and finite responses at every ROI vertex, leaving 4,930–9,000 images per subject and ROI. These images were absent from the 515-image fitting and reference-estimation set. The released response preprocessing, including session-wise normalization, was retained.

Eigenvector signs are arbitrary. We oriented the first fold's direction to correlate positively with the ROI-mean response on its training images. Other folds were aligned by the inner product of their coefficients expressed in the common, pre-standardization response coordinates. Each fold's projected scores were centered and scaled using its training-image score mean and standard deviation. We averaged these five standardized scores for image ranking, rather than averaging vectors expressed in different fold-specific coordinates. Zero denotes the training-image score mean, not a physiological baseline.

For each subject and hemisphere, we correlated all ten pairs of fold-specific scores on the same nonshared images. We averaged correlations over fold pairs and hemispheres before averaging subjects. Mean correlations were $0.972\ [0.964, 0.980]$ for V1v and $0.996\ [0.995, 0.997]$ for FFA (Figure 4A). Intervals throughout this section use 10,000 subject-bootstrap samples, retaining folds and hemispheres within subjects. This analysis measures stability of the response functions defined by different reference-image samples. Since each function is evaluated on the same measured responses, these correlations are not independent-repeat reliability estimates or noise ceilings.

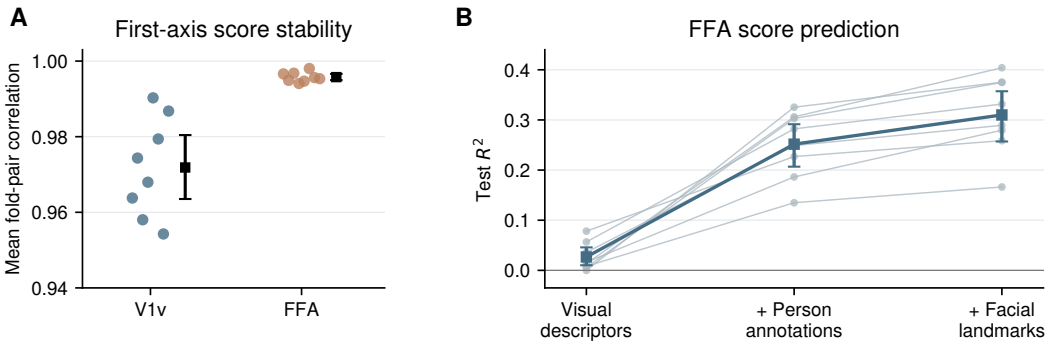

Figure 4: **First-axis stability and prediction from image properties.** (A) Correlations between scores from different reference folds on nonshared images. Colored points average fold pairs and hemispheres within each subject. Black squares show subject means and 95% bootstrap intervals. (B) Prediction of FFA scores on 1,000 nonshared images per subject. Light lines connect each subject's fold- and hemisphere-averaged test R^2 values. Dark squares show subject means and 95% bootstrap intervals. Feature sets are nested from left to right.

D.2 Prediction from image properties

For each fold, we fitted ridge regression from image descriptors to its first-axis scores using the 309 training images. Five-fold inner cross-validation selected the penalty from $\{10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4\}$. Feature standardization was fitted only on the corresponding training data. Evaluation used a fixed random sample of 1,000 eligible nonshared images per subject, shared across hemispheres and ROIs. We report $R^2 = 1 - \sum_i (y_i - \hat{y}_i)^2 / \sum_i (y_i - \bar{y}_{\text{test}})^2$, averaged over folds, hemispheres, and subjects in that order. No noise-ceiling normalization was applied.

The visual model used 31 descriptors of brightness and contrast, spatial-frequency and edge structure, texture, and color. It predicted V1v scores with $R^2 = 0.341\,[0.299, 0.380]$, compared with $0.027\,[0.010, 0.046]$ in FFA. For FFA, we added person and facial-landmark annotations from Common Objects in Context (COCO) (Lin et al., 2014), transformed to the actual NSD image crop. Person features measured presence, log-transformed count, and maximum cropped bounding-box area. Facial features measured presence, log-transformed count of people with a visible nose or eye inside the crop, and maximum visible-landmark span area. These annotations can miss small, unannotated, or animal faces, and landmark span is not face area.

Adding person annotations increased FFA test R^2 to $0.252\,[0.207, 0.291]$. Adding facial landmarks increased it further to $0.310\,[0.257, 0.357]$ (Figure 4B). The paired increment was $0.058\,[0.042, 0.076]$. Because inspection of subject 1's images motivated this analysis, we also excluded that subject. The increment remained $0.060\,[0.041, 0.079]$ across the other seven subjects, whose evaluation images were disjoint from subject 1's. This supports a face-related contribution to the FFA score beyond the measured visual and person features, consistent with the functional localizer used to define FFA. It does not establish a face-exclusive axis.

D.3 Image examples and comparison with other response summaries

Figure 5 shows subject 1's left V1v and FFA. For each ROI, we selected eight images with the highest positive scores, eight closest to zero, and eight with the most negative scores from all 9,000 eligible images. Selection used only the five-fold mean score, without manual selection by image content. V1v's positive examples often contain detailed textures and boundaries, while negative examples include large relatively uniform surfaces. FFA's positive examples include people and animals. These examples are descriptive and do not by themselves establish selectivity.

As an exploratory check of the V1v examples, we measured local brightness diversity in 1,000 nonshared images from subject 1. After resizing each image to 224×224 pixels, we divided grayscale

intensity into 16 fixed bins and computed Shannon entropy within each patch, then averaged patch entropies across the image. Correlations with the five-fold mean score, averaged over hemispheres, were 0.389, 0.426, and 0.444 for 4×4, 8×8, and 16×16 patch grids. This statistic describes local intensity distributions, not the spatial arrangement of edges. The grid comparison was exploratory, and these correlations leave substantial score variation unexplained.

We also compared the first TargetRef direction with the first signal-covariance eigenvector estimated by Generative Modeling of Signal and Noise (GSN), using the same fold and target coordinates. We aligned signs by their vector inner product, then correlated their scores on the same nonshared images. Their squared cosine similarities were 0.970 in V1v and 0.963 in FFA, and score correlations were approximately 0.995 in both (Table 7). These results indicate similar first reference directions across the two estimators. Principal component analysis (PCA) of the same 103 reference-image mean responses and the mean standardized ROI response also gave highly correlated scores. The image associations can, therefore, characterize this common leading response component. Agreement of the first direction does not imply equal weighting or agreement of later directions, which are compared in Appendix C.3.

Table 7: **First-axis image prediction and estimator comparisons.** Means and 95% subject-bootstrap intervals after averaging folds and hemispheres. GSN axis overlap is $(u_1^\top v_1)^2$. Score correlations use nonshared images. Facial-landmark increments use paired predictions and the same test-score variance denominator.

Quantity	V1v	FFA
Visual-descriptor test R^2	$0.341\,[0.299, 0.380]$	$0.027\,[0.010, 0.046]$
Facial-landmark increment in R^2	$-0.002\,[-0.004, 0.001]$	$0.058\,[0.042, 0.076]$
GSN first-axis overlap	$0.970\,[0.952, 0.983]$	$0.963\,[0.940, 0.980]$
GSN score correlation	$0.995\,[0.992, 0.998]$	$0.995\,[0.992, 0.998]$
PCA score correlation	$0.988\,[0.982, 0.994]$	$0.996\,[0.993, 0.998]$
ROI-mean score correlation	$0.969\,[0.964, 0.975]$	$0.970\,[0.962, 0.979]$



Figure 5: **Images ranked by the first TargetRef direction in V1v and FFA.** Subject 1, left hemisphere. Each row contains eight images selected by high positive, near-zero, or low negative five-fold mean scores. Numbers below images give the one-based NSD image identifier and standardized score. Scores are sign-aligned and standardized using training images before averaging folds. The sign is a convention, and near zero refers to the training-image mean. All displayed images are outside the shared 515-image analysis set. Images are NSD crops of COCO photographs (Allen et al., 2022; Lin et al., 2014). These photographs are excluded from this paper's license. Individual credits and terms of use are listed in Appendix I.

E ROBUSTNESS OF MODEL COMPARISONS

The rank, principal component analysis (PCA), null, and trial-partition controls below use the same target coordinates as the main-text TargetRef. We average the five folds for each source–target combination, then average sources within each subject×hemisphere×ROI target block. Means and 95% bootstrap intervals weight the 112 target blocks equally. Accuracy-matched model comparisons retain the 29 fixed pairs and weight their 26 eligible target blocks equally. Target-reference rank and repeat sensitivity are reported in Appendix C.4. The measured-motion control is described in Section E.5.

E.1 SOURCE RANK, RESPONSE VARIANCE, AND NULL CONTROLS

Source-rank sensitivity. We varied source rank while keeping TargetRef fixed (Figure 6C–D). Normalized ProfileMean scores were higher on average for pretrained than random models under adaptive rank selection (0.835 vs. 0.601), fixed rank 2 (0.961 vs. 0.809), and fixed rank 3 (0.986 vs. 0.826). Each score uses the median ProfileMean across the other seven subjects for the same target and source-rank condition as its denominator. These comparisons include all model–target units rather than only the accuracy-matched subset.

Relation to target-response variance. We fitted principal component analysis (PCA) to the 309 outer-training image-mean responses in the same vertex scaling as TargetRef, retaining components in descending variance order. We then measured their full-spectrum coverage of TargetRef, estimated from the 103 held-out images (Figure 6E–F). Coverage was $0.392\,[0.372, 0.413]$ for one component and $0.662\,[0.639, 0.683]$ for two. Matching PCA rank to the rounded mean rank of the 40 directional reference fits gave coverage $0.749\,[0.739, 0.759]$. Thus, variance-dominant training-response directions accounted for much of the held-out reference weight. PCA orders directions by response variance, whereas TargetRef weights them through the average projectors of regularized cross-trial predictive fits. The remaining difference can reflect subspace orientation, dimensionality, and the images used for estimation. These coverage values do not identify additional reliable signal specific to TargetRef. This target-side comparison is separate from the source-side PCA prediction-accuracy control in Figure 2D.

Random-subspace and response-permutation controls. Reference-rank-matched random target subspaces had mean full-spectrum coverage $0.0041\,[0.0039, 0.0044]$. Permuting training-image correspondences and refitting brain-source mappings gave coverage $0.393\,[0.375, 0.410]$, compared with $0.830\,[0.821, 0.838]$ for the unpermuted adaptive fits. Each fold used 16 random-subspace draws and four permutations per source subject, averaged before target-block inference. Permutation preserves target-response covariance, so fitted subspaces can retain overlap with variance-dominant directions despite shuffled image correspondence. The PCA and adaptive brain-source comparisons also differ in mean rank (2.48 vs. 3.86). Their coverage difference is therefore descriptive rather than a rank-controlled comparison of the two constructions.

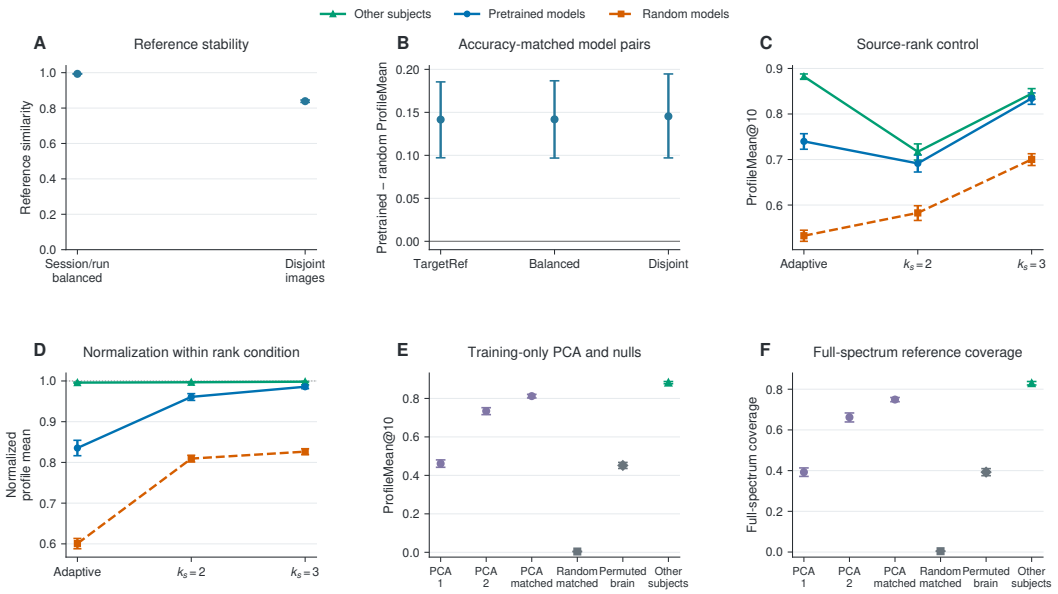

Figure 6: **Reference and source controls under TargetRef.** (A) Reference similarity after balancing trial partitions by session/run or using disjoint validation images. (B) Pretrained-minus-random ProfileMean differences for the same 29 accuracy-matched pairs across 26 equally weighted target blocks. (C) Source-rank sensitivity with TargetRef fixed. (D) The same ProfileMean scores normalized by the median across the other seven subjects for each target and rank condition. (E–F) ProfileMean and full-spectrum reference coverage for training-only target PCA, random target subspaces, response-permuted brain fits, and adaptive brain fits. Matched PCA and random-subspace ranks equal the rounded mean rank of the 40 directional reference fits. Adaptive brain fits select their own ranks. Random draws and permutations are averaged within folds. All panels show target-block means and 95% bootstrap intervals after fold averaging. Panels other than B use 112 blocks. Random models use seed 42.

E.2 Trial partitions and disjoint reference images

We rebuilt all 560 references using session/run-balanced trial partitions. Mean reference similarity with TargetRef was 0.9934. A separate reference estimated from the 103 validation images, retaining the original outer-training target scaling, had mean reference similarity 0.8390 (Figure 6A). The mean pretrained-minus-random ProfileMean difference for the fixed accuracy-matched pairs was 0.1419 with balanced partitions and 0.1454 with disjoint reference images, compared with 0.1416 for TargetRef (Figure 6B). The disjoint-image analysis changes the reference sample while preserving source fits, target coordinates, and accuracy-selected pairs. Its reference images were not used to fit or select the corresponding source subspaces or readouts.

We also rescore each saved source projector against a reference estimated from the 103 validation images in the same outer-training target coordinates (Figure 7). Across 896 model–target units after five-fold averaging, test-reference and validation-reference ProfileMean correlate at $r = 0.991$, with mean validation/test ratio 1.002. The corresponding correlation across 4,480 unaveraged model–fold entries is $r = 0.887$. These are descriptive reference-sensitivity checks with source fits held fixed.

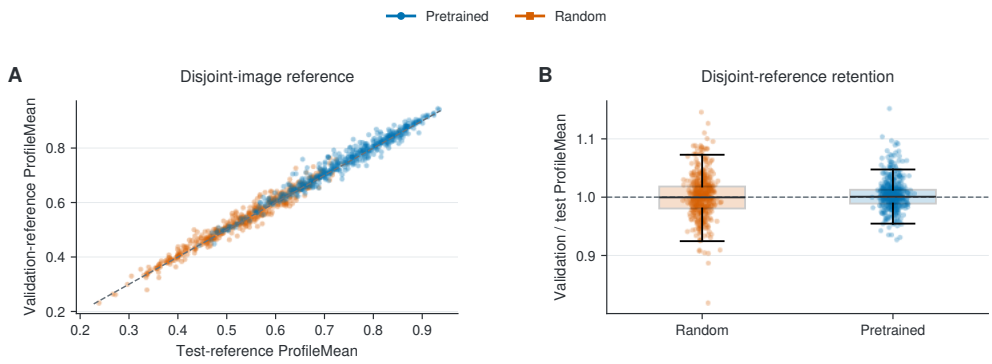

Figure 7: **Recovery profiles with disjoint reference images.** (A) ProfileMean using references estimated from the 103 test images and the separate 103 validation images. Both references retain the same outer-training target scaling and are evaluated against unchanged source subspaces. (B) Ratio of validation-reference to test-reference ProfileMean. Boxplots show medians, interquartile ranges, and 1.5-IQR whiskers. Both panels use five-fold averages for 896 model–target units. Random models use seed 42.

E.3 Profile length, accuracy matching, and directional prediction

The main text reports ProfileMean with $K = 10$ and defines near-equal-accuracy pairs using $|\Delta\text{accuracy}| \leq 0.01$. We varied these choices while keeping TargetRef and all source subspaces fixed (Table 8). Across all 448 same-architecture model pairs, the mean pretrained-minus-random difference changed little between $K = 5$, 10, and 20. The difference also remained positive for every combination of these K values and accuracy thresholds 0.005, 0.01, 0.02, and 0.03. Pairs were selected by five-fold mean accuracy alone. We averaged pair differences within each eligible target block before estimating the mean and its confidence interval.

The 0.01 tolerance bounds the magnitude of accuracy differences but does not balance their sign. We, therefore, inspected the signed accuracy and ProfileMean differences for all 29 fixed pairs (Figure 8). Pretrained models had higher accuracy in 21 pairs. In the other eight pairs, each from a different target block, six had higher pretrained ProfileMean and the mean ProfileMean difference was $+0.0791$. Higher ProfileMean thus also occurred without a pretrained advantage in observed prediction accuracy. This subgroup summary is descriptive and does not estimate an accuracy-adjusted effect of pretraining.

We also tested whether greater geometric coverage of a TargetRef direction accompanied more accurate prediction along that direction. We refitted the final ridge readout using each saved source basis, the original regularization parameter, and outer-training images only. The recomputed prediction accuracies differed from the saved values by at most 8.7×10^{-9}. We then measured $\text{corr}(\widehat{Y}_z u_j, Y_z u_j)$ on held-out images for the leading 20 directions, using the same outer-training target scaling as TargetRef. Directional coverage was positively associated with directional prediction in all three source groups (Table 8). These correlations describe associations across direction-level observations. Directions, folds, and sources are not treated as independent inferential samples.

Table 8: **Sensitivity and direction-wise prediction under TargetRef.** All controls use the fixed-coordinate 20-partition reference and saved source subspaces. Model differences are averaged over five folds, then eligible pairs within each target block. Means and 95% confidence intervals weight target blocks equally and use 10,000 paired bootstrap draws. Counts for the first two controls give pairs / eligible blocks. K sensitivity uses all same-architecture pretrained–random pairs. Threshold sensitivity uses $K = 10$. Directional correlations and direction counts are descriptive. The last column reports the difference between mean prediction correlations in the highest and lowest coverage quartiles after setting negative correlations to zero. Random models use seed 42.

Control	Setting	Count	Effect	95% CI	Additional summary
K sensitivity	$K = 5$	448 / 112	$+0.208$	$[0.194, 0.222]$	
	$K = 10$	448 / 112	$+0.207$	$[0.193, 0.222]$	
	$K = 20$	448 / 112	$+0.205$	$[0.190, 0.221]$	
Matched threshold	$\lvert\Delta\text{accuracy}\rvert \leq 0.005$	17 / 17	$+0.122$	$[0.064, 0.180]$	76% positive
	$\lvert\Delta\text{accuracy}\rvert \leq 0.010$	29 / 26	$+0.142$	$[0.097, 0.185]$	86% positive
	$\lvert\Delta\text{accuracy}\rvert \leq 0.020$	60 / 47	$+0.141$	$[0.109, 0.173]$	90% positive
	$\lvert\Delta\text{accuracy}\rvert \leq 0.030$	94 / 62	$+0.146$	$[0.119, 0.174]$	90% positive
Directional prediction	other subjects	78,400 dirs.	$r = 0.824$	–	high–low quartile $+0.382$
	pretrained	44,800 dirs.	$r = 0.726$	–	high–low quartile $+0.229$
	random	44,800 dirs.	$r = 0.628$	–	high–low quartile $+0.134$

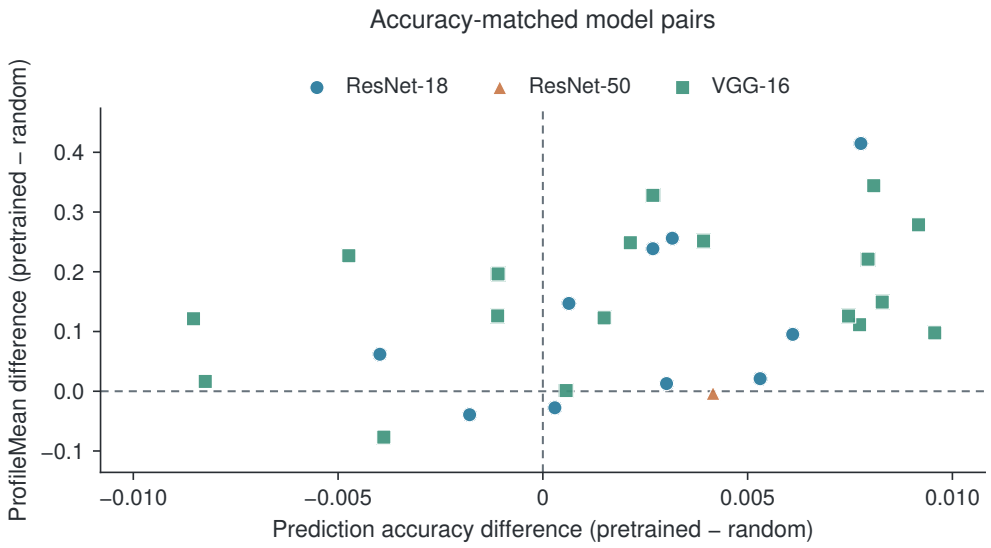

Figure 8: **Signed differences within the accuracy-matched model pairs.** Each point represents one of the 29 fixed same-architecture pairs across 26 target blocks. Both axes show pretrained-minus-random differences calculated from five-fold averages. ProfileMean uses TargetRef and $K = 10$. Dashed lines mark zero differences. Six of the eight pairs with nonpositive accuracy differences have positive ProfileMean differences. Colors and markers identify architectures. Random models use seed 42.

E.4 Random-initialization seeds

The main-text matched comparisons use shared random seed 42. We rescored the saved subspaces for four initialization seeds (0, 1, 2, and 42) against TargetRef to assess sensitivity to this choice (Table 9). The mean pretrained-minus-random ProfileMean difference across all architectures and target blocks was 0.171, compared with a standard deviation of 0.026 across the four aggregate random-seed means. Seed variability was greater for ResNet and VGG than for ViT-B/16. This aggregate comparison uses all model–target units, not the 29 accuracy-matched pairs. Figure 11 shows the corresponding profiles by ROI and architecture.

Table 9: **Random-initialization seed control under TargetRef.** Random controls use four shared initialization seeds (0, 1, 2, and 42) for each architecture. ProfileMean uses $K = 10$. We average five folds within each target unit and then give all 112 target units equal weight. Random-seed standard deviations (SDs) describe variation among four aggregate seed means, not confidence intervals. The overall row also averages the four architectures within each seed. The final column divides seed SD by the absolute pretrained-minus-random difference. Source subspaces are unchanged.

Architecture	Pretrained ProfileMean	Random seed mean	Random seed SD	Random min–max	Pretrained–random	SD / absolute difference
ResNet-18	0.717	0.620	0.058	0.546–0.668	0.098	0.594
ResNet-50	0.768	0.580	0.060	0.495–0.633	0.188	0.321
VGG-16	0.711	0.566	0.051	0.498–0.623	0.146	0.353
ViT-B/16	0.762	0.511	0.005	0.506–0.517	0.251	0.020
Overall	0.740	0.569	0.026	–	0.171	0.155

E.5 Measured-motion control

We formed 14 trial-level motion summaries. We projected the six rigid-body parameters, their six temporal derivatives, and Power framewise displacement (FD, using a 50-mm head radius) onto each trial's hemodynamic response regressor, and added maximum FD within 12 seconds of trial onset. For a trial regressor h_i and nuisance time series n_m, the projection is $z_{i,m} = h_i^\top n_m/(h_i^\top h_i)$. Decomposing each into an image-locked mean and within-image deviation produced a 28-column nuisance design. Session-specific coefficients were fitted by leave-one-run-out ridge regression with an intercept. Within each training-run set, an inner leave-one-run-out procedure selected the penalty from $\{0.1, 1, 10, 100\}$ by minimum mean squared error. Nuisance-feature means and scales were fitted only on the training runs of each split. TargetRef was rebuilt from residualized trials using unchanged trial identities, partitions, and outer-training target scaling, with no re-standardization after residualization.

All motion statistics below are *means across target blocks*, after fold averaging, with paired block-bootstrap intervals. Across 112 blocks, the residualized top-ten subspace retained 95.4% $[95.1\%, 95.6\%]$ of baseline reference weight, measured as $\mathrm{tr}(R_0 P_{\mathrm{res},1:10})/\,\mathrm{tr}(R_0)$, and baseline–residualized reference similarity was 0.9747 $[0.9731, 0.9761]$. For direction scores evaluated on image folds disjoint from reference estimation, the mean absolute correlation with FD over the leading three directions was 0.0582 $[0.0558, 0.0606]$. The mean within-block maximum was 0.0829 $[0.0786, 0.0876]$. We evaluated nuisance-only prediction relative to a session-intercept baseline using cross-validated coefficients of determination (R^2). All 336 block-by-direction values were negative after fold averaging.

Holding source subspaces and the accuracy-matched pairs fixed, the mean pretrained-minus-random ProfileMean difference over the 26 eligible blocks changed from 0.1416 $[0.0972, 0.1854]$ to 0.1508 $[0.1050, 0.1962]$. The measured-motion adjustment therefore did not reduce this difference. Trial-aligned eye data were too sparse, and physiological recordings lacked validated trial-aligned nuisance regressors, for equivalent full-sample controls. Repeat-stable eye movements, physiology, and other unmeasured nuisance remain possible contributors to the reference.

F Source-coordinate invariance and cross-region evaluation

F.1 Orthogonal rotations of source coordinates

For an already standardized source matrix X and orthogonal O, replacing X by XO leaves the isotropic ridge prediction $X(X^\top X + \lambda I)^{-1} X^\top Y$ unchanged. The corresponding reduced-rank prediction and target projector inherit this invariance when hyperparameters are fixed, subject to ambiguity if rank truncation cuts through an exactly tied singular subspace. No column-wise re-standardization is applied after rotation.

We performed two numerical checks across $4{,}480$ model–fold fits, covering four architectures in pretrained and randomly initialized states (random seed 42). First, we refitted the standardized reduced-rank ridge objective before and after orthogonal feature rotation, with rank and regulariza-

tion fixed. Target-projector overlap, $\mathrm{tr}(PP')/k$ for equal-rank projectors, exceeded $1 - 10^{-14}$ in every fit. The maximum relative prediction difference was 2.77×10^{-12}, and the maximum relative Gram-matrix difference was 7.99×10^{-16}. This standardized-objective check is separate from replaying the saved raw-coordinate readout.

We also re-expressed the saved fitted operators in rotated source coordinates without reselecting subspaces or hyperparameters. Let μ be the training-feature mean, D the diagonal matrix of training-feature standard deviations, and Q^X the saved source-side basis. For $X_{\mathrm{raw}} = \mu + X_z D$ and $X'_z = X_z O$, replacing Q^X by $O^\top Q^X$ preserves the target-induction coordinates $X_z Q^X$. Replacing DQ^X by $O^\top DQ^X$ and retaining the offset μQ^X preserves the raw-coordinate readout. Across the same fits, target-projector overlap again exceeded $1 - 10^{-14}$, and the maximum relative prediction difference was 5.28×10^{-14}. These projector and prediction checks do not depend on the reference estimator. Identical target projectors give identical recovery profiles for any fixed TargetRef.

Orthogonal coordinate invariance concerns reparameterization of a given source. It does not establish robustness to changing the allowed source-to-target mapping.

F.2 Cross-region alignment patterns on the same images

Alignment Pattern Analysis (APA) profiles cross-region relations against a structured human reference (Höfling et al., 2026). We evaluated an APA-inspired analysis on the same NSD folds, held-out images, previously extracted concatenated features, and fixed accuracy-matched pairs. This adaptation uses seven target ROIs and is not a literal replication of the original dataset and analysis. For each fold and hemisphere, we correlate the upper-triangular entries of correlation-distance representational dissimilarity matrices. The seven-component model pattern contains its similarity to each target ROI, divided by that target subject's mean same-ROI similarity to the other seven subjects. For each anchor ROI, the human pattern averages normalized cross-region similarities over the 42 ordered subject pairs among the other seven subjects. These cross-region pairs exclude the evaluated target subject, but their ROI-specific normalizers use all seven other subjects for each reference subject, including the evaluated target subject. Thus, exclusion of the evaluated subject applies to the cross-region pairs, not to every normalization term. Alignment Pattern Similarity (APS) is the Pearson correlation between the model and human patterns. The brain-to-brain benchmark used 115 accuracy-matched brain-source pairs on the same 26 eligible target blocks, with equal total weight per block.

The pretrained-minus-random APS difference was -0.0590 $[-0.1971, 0.0892]$. The corresponding TargetRef ProfileMean difference was 0.1416 $[0.0972, 0.1854]$. Both differences use target-block bootstrap intervals (Appendix B). Across 896 model–target units, APS and TargetRef ProfileMean correlated at $r = 0.316$ $[0.210, 0.414]$, with target-block cluster bootstrapping. At each metric's descriptive brain-to-brain threshold, absolute differences exceeded both thresholds for 15 pairs, only APA's for 6, and only TargetRef's for 8. The thresholds were 0.1447 for APS and 0.03355 for TargetRef. Each was the block-weighted 90th percentile of absolute differences among the same 115 accuracy-matched brain-source pairs across 26 target blocks. These absolute-difference classifications allow either direction of model preference and indicate complementary sensitivity. They are not tests of superiority. Inter-animal-transform work uses mappings between subjects to guide model–brain comparisons (Thobani et al., 2025). Here, we keep the mapping class fixed and use within-target differences between accuracy-matched brain sources as a benchmark for model differences.

G Results across ROIs and datasets

G.1 Brain-to-brain profiles

Brain-to-brain recovery under TargetRef is averaged over five folds, the seven other subjects per target block, and then subject–hemisphere blocks within each ROI. Across all target blocks, mean coverage is 0.951 at $k = 1$ and 0.853 at $k = 10$ (Figure 9).

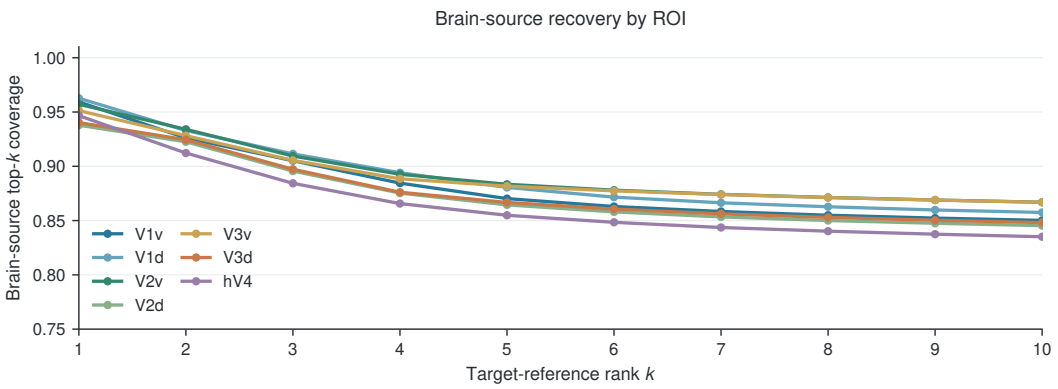

Figure 9: **Brain-to-brain recovery under TargetRef.** Mean profiles for the seven visual ROIs. Fold-level profiles are averaged over five folds, then the seven other subjects per target block, then 16 subject–hemisphere blocks per ROI. The sample contains 784 source–target units across 112 target blocks. Curves show arithmetic means.

G.2 Directional coverage

The main figures use top-k reference coverage to summarize recovery of a leading reference prefix. Figure 10 shows directional coverage and normalized TargetRef weights separately. Brain sources have the highest mean coverage of the first three directions, followed by pretrained and random models. Coverage generally decreases for later directions, which also carry less reference weight. Individual axes may be unstable within nearly tied eigenspaces, so we use these curves as descriptive diagnostics alongside the prefix profiles.

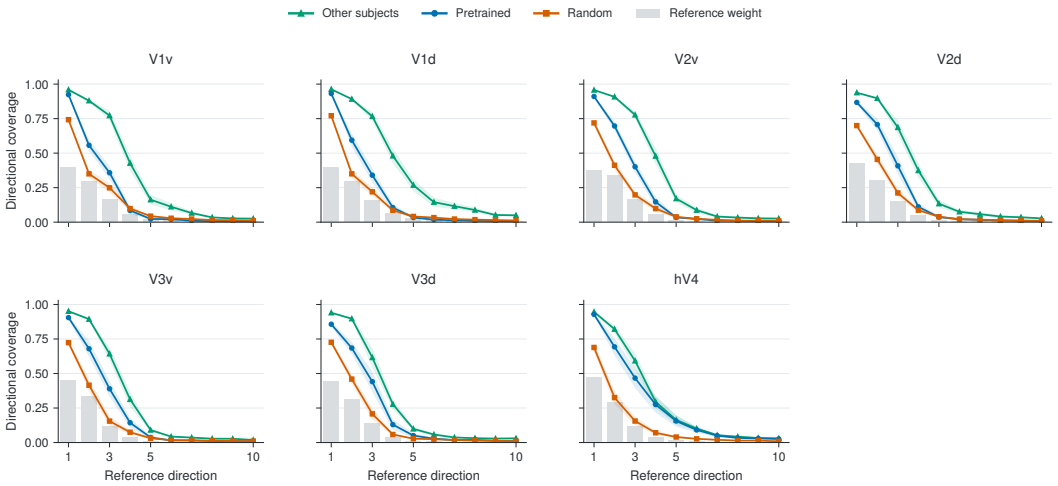

Figure 10: **Directional coverage under TargetRef.** Each panel shows directional coverage for the first ten directions in one visual ROI. We average five folds per source, then the seven other subjects or four architectures within each target block. Lines and bands show means and descriptive standard errors of the mean (SEMs) across the 16 subject–hemisphere blocks per ROI. Gray bars show the mean normalized reference weights. Random models use seed 42.

G.3 Model profiles by ROI and architecture

Figure 11 shows TargetRef recovery profiles by visual ROI and architecture for NSD-core-shared. Each panel compares pretrained and random models with the mean brain-to-brain profile for the same targets. Separating ROIs and architectures shows which differences persist across these conditions.

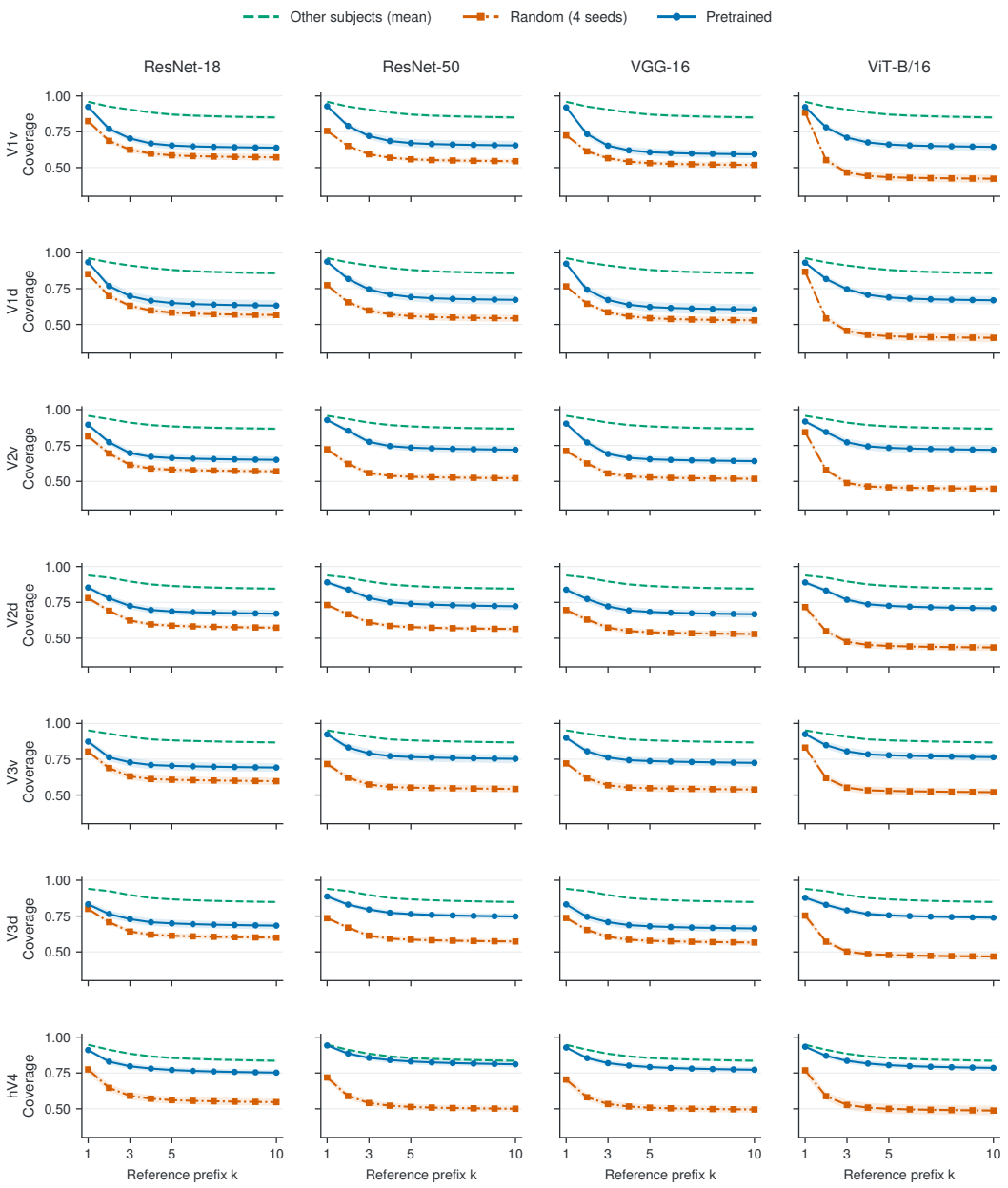

Figure 11: **NSD-core-shared model recovery profiles by ROI and architecture.** Each panel shows top-k reference coverage under TargetRef for one ROI and architecture. Coverage is a reference-weighted mean over the first k directions. Curves compare pretrained models, four-seed random controls, and other subjects. We first average five folds, then four random seeds or the seven other subjects within each target. Lines and shaded bands show means and standard errors of the mean (SEMs) across the 16 subject–hemisphere target blocks per ROI.

G.4 Dataset shift: NSD-synthetic

NSD-synthetic provides a dataset-shift check using the same subjects and different stimuli (Gifford et al., 2026). We constructed TargetRef in fixed outer-training coordinates for each dataset and rescored the saved source subspaces. NSD-synthetic uses 284 images with two to ten repeats, compared with 515 images and three selected trials for NSD-core-shared. The datasets also differ in response preprocessing, so this is not a controlled stimulus-only comparison.

The mean pretrained-minus-random ProfileMean difference across architectures was 0.171 on NSD-core-shared and 0.136 on NSD-synthetic after averaging the four random seeds. The difference was positive for each architecture in both datasets, but was only 0.010 for ViT-B/16 on NSD-synthetic. Thus, the separation depends on both the architecture and dataset. Figure 12 shows normalized ProfileMean alongside prediction accuracy.

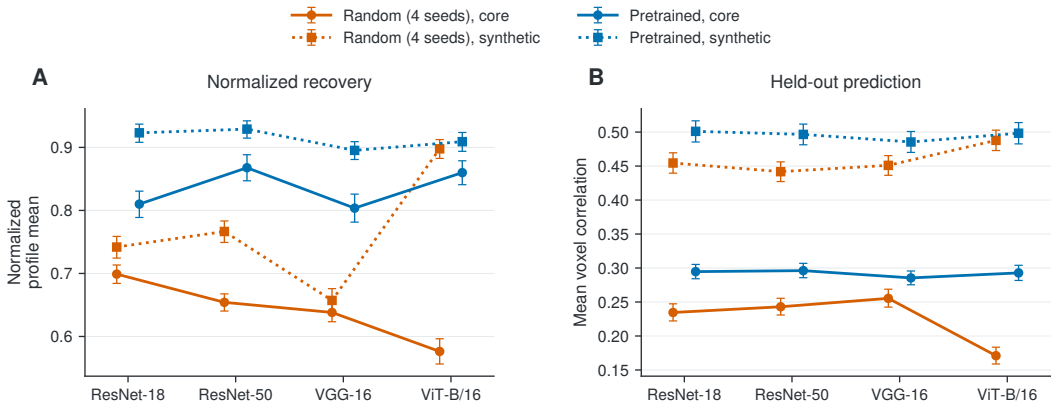

Figure 12: **Dataset-shift analysis.** (A) Normalized ProfileMean under TargetRef, using the median ProfileMean across the other seven subjects for the same target as the denominator. (B) Held-out prediction accuracy. Random controls average seeds 0, 1, 2, and 42 within each architecture and target. Points show means and 95% bootstrap intervals across 112 target blocks after fold averaging. Both datasets use fixed-coordinate 20-partition references, but differ in stimuli, repeat counts, and response preprocessing.

G.5 Higher-order visual and parietal ROIs

We applied TargetRef to five parcels from the Human Connectome Project Multi-Modal Parcellation version 1.0 (HCP-MMP1) (Glasser et al., 2016): posterior inferotemporal complex (PIT), fusiform face complex (FFC), posterior area TE2 (TE2p), intraparietal sulcus area 1 (IPS1), and intraparietal area 1 (IP1). These parcels were fixed as an exploratory set before inspecting reference or model-profile results. We mapped each subject's native HCP-MMP1 labels to the fsaverage surface used for the response data. We used the same subjects, folds, outer-training scaling, held-out sample size, repeats, and rank-selection rule as the early-visual comparison. Saved source subspaces and accuracy-only pair identities were retained. This analysis compares *pairs of pretrained architectures*. Because pair ordering is arbitrary, the outcome is the absolute ProfileMean difference. We first averaged each source's ProfileMean over folds, took absolute differences for eligible architecture pairs, then averaged these differences within each target block. The descriptive median effective rank was 2.748 across all 80 higher-parcel blocks and 4.289 across the 112 early-visual blocks.

Table 10 summarizes the accuracy-matched comparisons. After averaging ROIs within each of 16 subject $\times$ hemisphere units, the paired higher-minus-early difference was -0.0272 $[-0.0363, -0.0183]$. Diagnostic separation therefore did not increase in this exploratory ROI set with the available images and repeats. The result does not isolate reference-estimation precision, source recoverability, stimulus coverage, or brain-to-brain variability, and does not establish intrinsically lower dimensionality in association cortex.

We defined the functional fusiform face area (FFA) as the union of the NSD `floc-faces` labels `FFA-1` and `FFA-2`, mapped from each subject's native surface to fsaverage. This separate analysis gave a median TargetRef effective rank of 2.660. Across 16 eligible target blocks, the mean absolute ProfileMean difference between pretrained architectures was 0.0146 $[0.0074, 0.0241]$, with a ratio of 0.609 $[0.309, 1.016]$ to the brain-to-brain threshold. The interval is conditional on that threshold.

Table 10: Exploratory comparison of accuracy-matched pretrained architectures under TargetRef. Ratios divide each absolute model-pair ProfileMean difference by the ROI-specific 90th percentile of absolute ProfileMean differences between accuracy-matched brain sources. Bootstrap intervals hold this percentile fixed.

Quantity	Higher-order parcels	Early-visual ROIs
Eligible target blocks	76	110
Mean absolute ProfileMean difference	0.0248 [0.0200, 0.0301]	0.0519 [0.0457, 0.0584]
Ratio to brain-to-brain threshold	0.372 [0.294, 0.458]	1.558 [1.332, 1.808]

G.6 Layer and pretraining-objective comparisons

Figure 13 shows how recovery varies across layers and training objectives. These comparisons use the same TargetRef and normalization against other subjects and are descriptive. In the ResNet-50 family, we compare ImageNet supervision, self-distillation with no labels (DINO) (Caron et al., 2021), Contrastive Language–Image Pre-training (CLIP) (Radford et al., 2021), and an L2-robust checkpoint from the MadryLab `robustness` library (Engstrom et al., 2019). Adversarially robust convolutional neural networks have previously been reported to better match several macaque V1 response properties (Kong et al., 2022). The ViT comparison also includes DINOv3 (Siméoni et al., 2025) and CLIP ViT-B/16 (panel E). The figure summarizes all target blocks without accuracy matching.

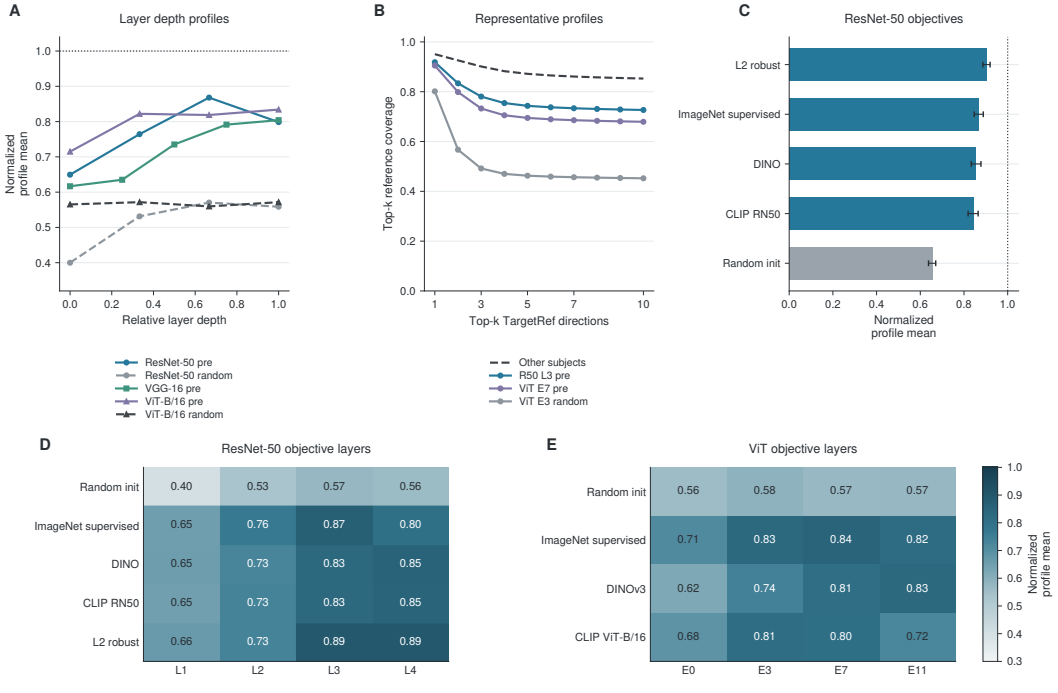

Figure 13: **Layer-wise and objective diagnostics under TargetRef.** (A) Normalized ProfileMean across normalized layer depth for representative pretrained and random sources. (B) Profiles for pretrained ResNet-50 layer 3, pretrained ViT block 7, and random ViT block 3, with the mean profile from the other seven subjects. These examples are held fixed across reference comparisons. (C) ResNet-50 objective comparison, with 95% target-block bootstrap intervals. (D–E) Normalized ProfileMean by objective and layer for ResNet-50 and ViT on a shared color scale. We average folds, divide each source's ProfileMean by the median across the other seven subjects, and then average the 112 target blocks with equal weight. Layer and objective comparisons are descriptive. Random models use seed 42.

Accuracy-tolerance sensitivity. We separately varied the maximum allowed accuracy difference for the pooled stage 1–4 representations, retaining the source subspaces and TargetRef (Table 11). The 0.01 criterion yielded few eligible pairs. At tolerances of 0.02 or greater, all four conditions had positive mean ProfileMean differences. Wider tolerances also admitted a pretrained advantage in prediction accuracy, so these results do not separate the profile difference from differences in accuracy. The main 29-pair analysis retains its original 0.01 criterion.

Table 11: **ResNet-50-family ProfileMean differences across accuracy tolerances.** Entries give the mean pretrained-minus-random ProfileMean difference, with the number of eligible pairs in parentheses. Each condition has 112 candidate subject–hemisphere–ROI blocks and at most one pair per block. Pairs are selected by five-fold mean accuracy alone. Eligible blocks receive equal weight. The random baseline is standard ResNet-50 with seed 42. CLIP RN50 uses a modified architecture, so its comparison is not architecture-matched. A dash indicates no eligible pairs. These point estimates are descriptive, particularly for small subsets.

Maximum absolute accuracy difference	ImageNet supervised	DINO	L2-robust	CLIP RN50
0.01	−0.003 (1)	– (0)	+0.101 (1)	+0.058 (4)
0.02	+0.077 (6)	+0.088 (6)	+0.124 (3)	+0.052 (10)
0.03	+0.088 (13)	+0.113 (15)	+0.147 (6)	+0.092 (19)
0.05	+0.122 (36)	+0.113 (40)	+0.162 (15)	+0.098 (42)
0.10	+0.163 (90)	+0.153 (90)	+0.194 (81)	+0.138 (88)

H IMPLEMENTATION AND REPRODUCIBILITY

Running `python run.py –output results` in the accompanying code recomputes numerical summaries and confidence intervals from the included fold-level or target-block measurements. The README gives the refitting command for separately supplied, preprocessed brain responses and model features. The `config/` directory fixes folds, image IDs, trial triplets, model pairs, and feature definitions. `INPUTS.md` defines input and intermediate arrays. Raw fMRI data, response and feature arrays, stimulus images, and third-party model checkpoints are not included. Credits for the stimulus examples reproduced in Figure 5 are listed in Appendix I.

The analyses use NSD, NSD-synthetic, GLMsingle response estimates, and model definitions and weights from the PyTorch vision library (torchvision), PyTorch Image Models (timm), and the checkpoint providers listed in Table 12. GSN and neuropop provide the signal-covariance and cvPCA estimators. The table summarizes these resources, their roles, access requirements, and licenses or terms of use. We provide the analysis code as supplementary material under the MIT License. Third-party resources remain subject to their original terms.

Table 12: **Datasets, software, and model checkpoints.** Data and checkpoints are obtained from their original providers under the corresponding terms.

Resource	Role	Access	License or terms
NSD responses and images (Allen et al., 2022)	Main fMRI evaluation and model feature extraction	Local dataset obtained through NSD data access.	NSD Data Access Agreement and Terms and Conditions.
GLMsingle response estimates (Prince et al., 2022)	Trial-level beta estimates for repeated-response views	Precomputed arrays in the documented format.	Response estimates follow NSD or NSD-synthetic data terms. GLMsingle code: BSD 3-Clause.
GSN v1.2.0 (Kay et al., 2025)	Signal and noise covariance estimation	Official Python implementation.	BSD 3-Clause.
neuropop v1.0.1 (Stringer et al., 2019)	Cross-validated PCA spectrum estimation	Official Python implementation.	GPL v3.
NSD-synthetic (Gifford et al., 2026)	Dataset-shift validation	Local dataset obtained through NSD data access.	Original NSD/NSD-synthetic data terms.
ImageNet images	Pretraining data for ImageNet-trained checkpoints	Not required as evaluation inputs.	ImageNet Terms of Access, including non-commercial research and educational use. Image copyrights remain with their owners.
torchvision/timm architectures and weights	Model feature extraction	Software packages and public pretrained weights. Random controls are initialized locally.	torchvision code: BSD 3-Clause. timm code: Apache 2.0. Pretrained weights may have additional dataset or provider terms, including ImageNet terms.
ResNet-50 checkpoints	Comparison of ImageNet supervision, DINO, CLIP, and L2-robust training (Radford et al., 2021; Caron et al., 2021; Engstrom et al., 2019)	Public checkpoints from the original providers.	OpenAI CLIP: MIT. DINO: Apache 2.0. MadryLab robustness and robust ImageNet models: MIT.

RRR and recovery-profile computations run on CPUs once model features are available. For one 710-vertex V1v target, five folds, and three sources (pretrained/random ResNet-18 and another subject), a single-worker CPU run took 202 s with a sampled peak process-tree RSS of 0.55 GiB on an Intel Xeon Silver 4314. This includes TargetRef, GSN, cvPCA, source fitting, and summaries from prepared arrays. Feature extraction and the official GSN implementation can use GPUs. Runtime and memory depend on ROI size, feature dimensionality, and available hardware.

Collection of the NSD dataset was supported by NSF IIS-1822683 and NSF IIS-1822929.

I IMAGE CREDITS

Figure 5 reproduces cropped stimulus images from the Natural Scenes Dataset (NSD), derived from Common Objects in Context (COCO) photographs. We used the existing NSD crops and resized them for display. The photographs are excluded from this paper's license and remain subject to the individual terms listed below.

The tables follow the image order in the figure. Each title links to its Flickr source, and each license label links to the corresponding terms. CC licenses or public-domain dedications are taken from the source pages where available and otherwise from the COCO records. Creator names and titles were retrieved from the source pages where available.

*Creator and title could not be retrieved from the linked public page.

V1v

NSD ID	Creator and photograph title	License
Positive		
10477	Chris Phan. Sandy River with train in background.	CC BY 2.0
33265	Eli Duke. Attleson Farm: Pizza for Lunch.	CC BY-SA 2.0
64747	Design for Health by Ann Forsyth. Inverell, Australia PC270107.	CC BY 2.0
53446	steve p2008. next doors cat.	CC BY 2.0
70827	KENT BERGMAN. Horse-2013-45.	CC BY 2.0
3478	Julie Pi. At the market.	CC BY-NC 2.0
14888*	Flickr photograph 3787184751	CC BY-NC-ND 2.0
50923*	Flickr photograph 9423457108	CC BY-NC-SA 2.0
Near zero		
8859	Bruna Ferrara;. :P.	CC BY-NC-ND 2.0
17109	Nehama Verter. Max was surprisingly okay with being brushed.	CC BY-NC-SA 2.0
57072*	Flickr photograph 5507320891	CC BY-NC 2.0
61981*	Flickr photograph 2169562836	CC BY-ND 2.0
44870	Neil Howard. Longis Boat - Alderney.	CC BY-NC 2.0
70585	Smudge 9000. Grey Phalarope mating.	CC BY-SA 2.0
33867	clydesbale. rockies.	CC BY-NC 2.0
47204	Alan Myers. Umbrella.	CC BY 2.0
Negative		
42394*	Flickr photograph 6764231277	CC BY-NC-ND 2.0
28430	Dmitry Shakin. Skyline.	CC BY-NC-ND 2.0 (COCO record)
11148	Steve Cox. 8th Street signs.	CC BY-NC-SA 2.0
54563	summerreader. Dining room.	CC BY-NC-SA 2.0
29512	Macomb Paynes. Mississippi River Drive.	CC BY-NC-SA 2.0
69028	saimad. Adding to the motorcycle collection.	CC BY-SA 2.0
52345	Photo Monkey. Hostel colours.	CC BY 2.0
51842	George Thomas. Finch over Water.	CC BY-NC-ND 2.0

FFA

NSD ID	Creator and photograph title	License
Positive		
66304*	Flickr photograph 4237884823	CC BY-NC-ND 2.0
2673	Dark Dwarf. Cutting the Cake.	CC BY-ND 2.0
63170*	Flickr photograph 8031325821	CC BY 2.0
39409	William Kunz. They Have Ways of Making You Talk.	CC BY-NC-ND 2.0
58157	Maia C. Dive!	CC BY-NC-ND 2.0
319	Jessie Romaneix Gosselin. Porté couleur.	CC BY-NC-ND 2.0
35971	Stephen Cummings. Picking out a tie.	CC BY 2.0
15686	Buster Benson. 8:36pm To brush your teeth or not to brush your teeth #3yearoldproblems.	CC BY-SA 2.0
Near zero		
35198	Alan Levine. And Then There Was One (Laptop).	CC0 1.0
32047	Arno Meintjes. African elephant.	CC BY-NC-SA 2.0
22712*	Flickr photograph 1589440843	CC BY-NC-ND 2.0
11859	Kathrin Rieger. Teddy.	CC BY 2.0
48741	Rafal Kiermacz. Wall Street sign in New York, USA.	CC BY-NC-SA 2.0
46648	Bryan Costin. Yes, Mr. President. I'll get right on it.	CC BY-NC-SA 2.0
53067	Keith Williamson. Next door's cat Project 365(3) Day 223.	CC BY 2.0
66526	David. 8:36pm Watching more Giants baseball.	CC BY-NC-SA 2.0
Negative		
33657	momoko-tomoko. DSC_0071.	CC BY-NC 2.0
10495	Bernard Blanc. Le bateau camouflé, village des Jacquets, Lège-Cap Ferret, Gironde, Aquitaine, France.	CC BY-NC-SA 2.0
29600*	Flickr photograph 2662547557	CC BY-SA 2.0
16709	Natalia Wilson. Maxx preparing pizza.	CC BY-SA 2.0
72535	Tom Britt. Warner-Kitchen.	CC BY 2.0
36357	dereksemmler. Gemini Gas Range - Top Oven Open.	CC BY-SA 2.0
60364	Agriturismo Le Docce - Tuscany. Apartment Noce-Agriturismo LeDocce.	CC BY 2.0
60545	William Warby. Bathroom.	CC BY 2.0