DO FMRI ENCODING MODELS PRESERVE CROSS-SUBJECT TRANSFER? DIAGNOSING AND REDUCING THE BRAIN–MODEL TRANSFER GAP

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-subject transfer is important for fMRI encoding models, but whether model-predicted responses preserve the transfer capability of measured brain responses remains unclear. We propose brain-referenced transfer evaluation, which compares transfer from measured source-subject responses with transfer from model predictions under matched target adaptation. On NSD, model predictions from 4 backbones $\times$ 3 predictors transfer less effectively than measured responses, showing that training models to predict source responses does not ensure full transfer preservation. We probe this gap in the primary DINOv2–TBEn configuration, finding that predicted responses concentrate variance in fewer principal components (PCs) than measured responses and that leading PCs of measured responses support more transfer than the same number of model PCs. Correcting model predictions with a mapping learned from measured source responses narrows the transfer gap. These findings motivate Population Subspace Supervision (PSS), which upweights training errors in the leading source-population subspace. PSS closes 39.5% of the brain–model transfer gap, improves mean transfer across all 12 NSD configurations, and improves mean correlation in 11 of 12 BOLD5000 configurations. Together, these results highlight cross-subject transfer as a design objective for fMRI encoding model training alongside source-response prediction.

1 INTRODUCTION

Modeling and predicting human brain activity is a central goal in cognitive neuroscience. Encoding models predict stimulus-evoked fMRI responses with linear (Kay et al., 2008; Huth et al., 2012; Güçlü & Van Gerven, 2015; Huth et al., 2016; Antonello et al., 2023) and nonlinear methods (Wen et al., 2018b; Adeli et al., 2026; Beliy et al., 2024), enabling increasingly general-purpose fMRI foundation models (Beliy et al., 2024; d'Ascoli et al., 2026; Moussa & Toneva, 2026). Yet accurate models for new individuals require large amounts of costly fMRI data, motivating cross-subject knowledge transfer (Wen et al., 2018a; Khosla et al., 2020). Recent methods improve cross-subject transfer via multi-subject learning, meta-learning, subject conditioning, anatomical standardization, and shared–individual modeling (Beliy et al., 2024; Yu et al., 2026b; d'Ascoli et al., 2026; Chen et al., 2026; Moussa & Toneva, 2026). These advances increasingly enable prediction of a new subject's responses from information learned from others, raising an important question for individualized modeling: whether predictions retain the transfer capability of measured source responses.

However, existing methods primarily demonstrate transfer through improved target-subject prediction, leaving unassessed how much transfer capability model predictions retain relative to measured responses. Recent brain–model evaluations emphasize relational structure in real brains, making measured source responses a natural transfer reference (Thobani et al., 2025; Höfling et al., 2026). Moreover, model-predicted brain responses have helped train other individuals' encoding models, making their transfer capability directly relevant to downstream use (Wasserman et al., 2024).

To assess how well model-predicted responses retain the cross-subject transfer capability of measured source responses, we propose brain-referenced transfer evaluation, using measured source-subject fMRI as an empirical brain reference. We compare the ability of model-predicted and measured source responses to predict target responses under the same target support and query sets and

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

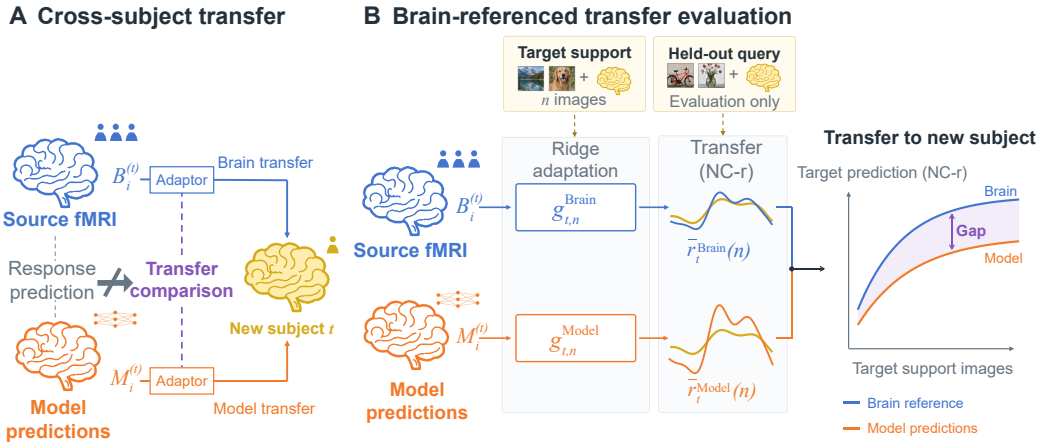


Figure 1: **Response prediction and brain-referenced transfer evaluation. A,** Source-response prediction and cross-subject transfer after target adaptation are distinct: training models to predict source fMRI does not by itself imply that the predicted responses will support equally effective transfer to a new subject (yellow). Our evaluation aims to compare transfer from measured source-population responses (Brain, blue) with transfer from model predictions (Model, orange). **B,** Brain (blue) and Model (orange) responses are mapped to the target subject (yellow) by separate Ridge adaptors fitted on the same n-image support set and evaluated on matched held-out queries. The schematic curves illustrate the Brain-Model transfer gap across target support sizes.

the same adaptor family (Fig. 1B). Across all evaluated NSD backbone–predictor configurations, baseline model predictions achieve lower mean transfer than the brain reference. These results separate source-response prediction from transfer preservation: training models to predict source responses alone does not ensure retaining their predictive value for new subjects, exposing an important limitation of current prediction-based training (Fig. 1A).

Guided by work linking representational geometry to prediction and generalization (Chung & Abbott, 2021; Elmoznino & Bonner, 2024; Liu et al., 2026; Yu et al., 2026a), we analyze predicted-response geometry and its relation to transfer. We find that model predictions exhibit effective-rank collapse, concentrating response variance in a few directions, while additional leading principal components (PCs) from source fMRI support more transfer than equally sized sets of Model PCs. A coordinate map learned on source training data improves transfer from the existing Model outputs. An oracle intervention further shows that replacing coordinates in the leading Brain-PC subspace reaches the source-fMRI reference while preserving the remaining Model response. These findings motivate **Population Subspace Supervision (PSS)**, which more heavily penalizes prediction errors in these population coordinates during training. In the primary configuration, PSS improves response geometry and coordinate recovery; it also improves cross-subject transfer across all evaluated NSD configurations and the target-support range, with additional BOLD5000 validation.

Our main contributions are:

- **Evaluation framework:** We propose brain-referenced transfer evaluation and show across all evaluated NSD configurations that model predictions transfer less effectively than measured source fMRI; training models to predict source responses alone therefore does not ensure full transfer preservation.
- **Diagnostic analysis:** We show that predicted responses concentrate variance in fewer directions, while leading Brain PCs support greater transfer than equally sized sets of Model PCs. Learned coordinate correction reduces the gap, and oracle replacement within the leading Brain-PC subspace reaches the brain reference.
- **Training improvement:** We propose Population Subspace Supervision (PSS), which strengthens population-coordinate supervision and improves mean transfer across all evaluated NSD configurations, with additional validation on BOLD5000.

2 Related Work

fMRI encoding and cross-subject transfer. fMRI encoding models predict stimulus-evoked responses from visual and semantic features (Kay et al., 2008; Huth et al., 2012; 2016; Wen et al., 2018b; Adeli et al., 2026). Transfer learning and shared encoders reduce individual data needs by reusing information across subjects (Wen et al., 2018a; Khosla et al., 2020). Recent approaches use distinct adaptation mechanisms: UBE combines multi-subject learning with voxel embeddings (Beliy et al., 2024); BraInCoRL uses meta-learned in-context adaptation (Yu et al., 2026b); and NRF uses standardized anatomical coordinates (Chen et al., 2026). Population modeling also covers multimodal responses in TRIBE v2 (d'Ascoli et al., 2026) and zero- and few-shot speech encoding in RABBiT (Moussa & Toneva, 2026). We complement these approaches by comparing transfer from predicted versus source-fMRI responses and using the resulting gap to guide source-model training.

Brain-referenced evaluation of neural models. Brain–model comparisons commonly use linear predictivity and representational similarity (Yamins et al., 2014; Kriegeskorte et al., 2008), often combined across neural and behavioral benchmarks (Schrimpf et al., 2018; 2020). Because conclusions and rankings can depend strongly on the alignment measure while diverse models achieve similar scores (Soni et al., 2024; Ahlert et al., 2024; Conwell et al., 2024; Wu et al., 2025; Klabunde et al., 2025), recent work seeks stronger criteria for brain alignment. The brain hierarchy score measures hierarchical correspondence between model layers and cortical regions (Nonaka et al., 2021), while the NeuroAI Turing Test benchmarks model–brain similarity against brain–brain variability (Feather et al., 2025). IATC constrains model–brain comparison to mappings sufficient for inter-individual alignment (Thobani et al., 2025), whereas APA compares region-wise model–brain alignment patterns with their brain–brain counterparts (Höfling et al., 2026). Following these efforts, we use source-subject fMRI responses as an empirical reference to diagnose and improve cross-subject transfer.

Shared neural structure and representational geometry. Cross-subject analyses have long sought population-level neural structure shared across individuals. Despite idiosyncratic functional topographies, hyperalignment and searchlight hyperalignment estimate common representational spaces (Haxby et al., 2011; Guntupalli et al., 2016), while Shared Response Models and connectivity extensions factor multi-subject responses into shared representations and subject-specific mappings (Chen et al., 2015; Nastase et al., 2020). Intersubject analyses complement these approaches by characterizing stimulus-driven responses shared across brains (Nastase et al., 2019). More broadly, representational and neural-population geometry characterize information through the organization and dimensionality of population activity (Kriegeskorte & Kievit, 2013; Kriegeskorte & Wei, 2021; Chung & Abbott, 2021). Recent work links internal-representation geometry—including trajectory curvature, latent or intrinsic dimensionality, and shared representational dimensions—to predictive performance, generalization, and brain alignment (Hosseini & Fedorenko, 2023; Elmoznino & Bonner, 2024; Liu et al., 2026; Mahner et al., 2025; Chen & Bonner, 2025; Yu et al., 2026a). We complement work on shared brain spaces and internal model geometry by examining predicted-fMRI geometry and its relation to cross-subject transfer.

3 Methods

3.1 Experimental setup

We used the Natural Scenes Dataset (NSD; Allen et al., 2022) as the primary experimental setting, focusing on the four participants who completed all 40 scanning sessions and the Shared1000 images viewed by all four. We performed leave-one-subject-out (LOSO) evaluation, treating one participant as the target and the remaining three as sources. Within each LOSO split, we used five-fold cross-validation with joint image and session holdouts. Images and sessions received separate fold labels: source-model training used trials with neither held-out label, and the target query included only trials for which both labels matched the held-out fold. Target adaptation used the target participant's trials from n distinct images in the outer-training pool, excluding held-out images and sessions. The predictor is trained on source responses, the target adaptor is fitted on target support, and transfer is evaluated on the held-out target query (Fig. 1B, Appendix A).

Each encoding model paired a frozen visual backbone H with an fMRI predictor f_{θ_t} trained on source responses in LOSO split t; $H(x_i)$ denotes the feature vector for image x_i. We evaluated 12 configurations combining four backbones—DINOv2 (Oquab et al., 2023), AugReg (Steiner et al., 2021), LAION2B-pretrained CLIP (Radford et al., 2021; Cherti et al., 2023), and ResNet-50 (He et al., 2016)—with three predictors: TBEn (Adeli et al., 2026), Ridge, and UBE (Beliy et al., 2024). DINOv2–TBEn was primary.

We also evaluated PSS on four BOLD5000 participants and their 2,952 shared images, using raw Pearson correlation for transfer (Chang et al., 2019) (Appendix F).

For query trials Q, let $r_v = \mathrm{corr}_{i \in Q}(y_{iv}, \hat{y}_{iv})$ denote raw Pearson correlation at vertex v. NSD transfer was evaluated using

$$r = \frac{1}{V} \sum_{v=1}^{V} r_v, \qquad \text{NC-}r = \frac{r}{V^{-1} \sum_{v=1}^{V} \mathrm{NC}_v}, \tag{1}$$

Here y_{iv} and $\hat{y}_{iv}$ are measured and predicted target responses, and NC_v is the target's single-trial correlation noise ceiling at vertex v. Metrics were computed separately for each hemisphere, with V denoting its evaluated vertices, and then averaged within each target and support size. Noise-ceiling estimation and metric aggregation are detailed in Appendix A.3.

3.2 Brain-referenced transfer evaluation

We compare predicted and source-fMRI responses as inputs for adaptation to a new subject (Fig. 1B). For target t, let $\mathcal{S}_t$ denote the source subjects. We define the Brain and Model responses to image i as

$$B_i^{(t)} = \frac{1}{|\mathcal{S}_t|} \sum_{s \in \mathcal{S}_t} \bar{y}_{s,i}, \qquad M_i^{(t)} = f_{\theta_t}(H(x_i)), \tag{2}$$

Here $\bar{y}_{s,i}$ is source subject s's response averaged over repeats. Brain denotes the measured source-population response, and $M_i^{(t)}$ the model prediction in the same cortical space. For UBE, Model averages predictions across the three sources.

Both paths use identical target support and query trials. Let $g_{t,n}^{c}$ denote the adaptor for target t, support size n, and condition c. It is fitted from condition c's representation to target responses on support; Brain uses measured source responses for both support and query images. Query responses provide evaluation scores. Plotted NSD analyses use Ridge with fixed $\alpha = 10$; the additional primary comparison selects regularization within target support (Appendix A.4), and nonlinear adaptors are evaluated in Appendix E.2.

Let $\bar{r}_t^{\mathrm{Brain}}(n)$ and $\bar{r}_t^{\mathrm{Model}}(n)$ denote Eq. (1) averaged over folds, hemispheres, and repeated factors within target t. We define the **n-shot relative transfer score** as the Model-to-Brain ratio at the same support size:

$$b_t(n) = \frac{\bar{r}_t^{\mathrm{Model}}(n)}{\bar{r}_t^{\mathrm{Brain}}(n)}. \tag{3}$$

This ratio expresses Model transfer relative to the brain reference: 1 denotes equal performance, and values below or above 1 indicate lower or higher transfer.

We summarize transfer across support sizes using the **log-shot relative transfer score** b_t. For each target, we divide the Model transfer-curve area by the Brain transfer-curve area on the log-shot axis, then average the ratios across targets:

$$b_t = \frac{\mathrm{AUC}_{\log n}\left(\bar{r}_t^{\mathrm{Model}}(n)\right)}{\mathrm{AUC}_{\log n}\left(\bar{r}_t^{\mathrm{Brain}}(n)\right)}, \qquad \bar{b} = \frac{1}{T} \sum_{t=1}^{T} b_t, \tag{4}$$

where $T = 4$. All transfer analyses use eight support sizes: 5, 10, 20, 50, 100, 200, 500, and full support. Other conditions are scored by replacing the Model curve in the numerator; the denominator remains the transfer AUC from the full Brain response, using all response dimensions within the evaluated cortical space. Full support denotes the maximum available target support set. We report target-prediction NC-r and paired absolute differences alongside relative transfer and gap closure. Aggregation and uncertainty procedures are detailed in Appendix B.

3.3 MECHANISM ANALYSIS

We tested whether Model responses concentrate variance, additional Brain PCs support transfer, and Brain-PC coordinate correction restores transfer; these diagnostic analyses leave baseline training unchanged.

Response geometry. Within each LOSO split, we separately fitted PCA to Brain and Model responses on source-training images. This yielded means $\mu_{\mathrm{B}}, \mu_{\mathrm{M}}$ and leading-k bases $U_k^{\mathrm{B}}, U_k^{\mathrm{M}} \in \mathbb{R}^{k\times V}$, whose rows are orthonormal PC directions. For a response a, its coordinates in the Brain and Model bases are

$$z_k^{\mathrm{B}}(a) = (a-\mu_{\mathrm{B}})(U_k^{\mathrm{B}})^\top, \qquad z_k^{\mathrm{M}}(a) = (a-\mu_{\mathrm{M}})(U_k^{\mathrm{M}})^\top. \tag{5}$$

Thus, $z_k^{\mathrm{B}}(B_i)$ and $z_k^{\mathrm{M}}(M_i)$ give coordinates in each response's own basis, whereas $z_k^{\mathrm{B}}(M_i)$ projects a Model response onto the Brain basis.

The means and bases were fixed before target adaptation. We characterized variance concentration using cumulative explained variance and effective rank (ER). For the leading $K = 64$ singular values of the centered Brain and Model response matrices, define

$$p_j^c = \frac{(s_j^c)^2}{\sum_{\ell=1}^{K}(s_\ell^c)^2}, \qquad \mathrm{ER}_K^c = \exp\left(-\sum_{j=1}^{K} p_j^c \log p_j^c\right), \quad c \in \{\mathrm{B},\mathrm{M}\}, \tag{6}$$

Cumulative explained variance is normalized by total response variance, whereas ER is normalized within the leading 64 PCs (Appendix C.2). A spectrum-replacement control, detailed in Appendix C.3, tested whether redistributing variance along existing directions altered transfer.

Transfer from leading PCs. We fitted target adaptors using the leading k Brain coordinates, $z_k^{\mathrm{B}}(B_i)$, or Model coordinates, $z_k^{\mathrm{M}}(M_i)$, for $k \in \{1,2,3,4,8,16,32,64\}$. Each condition used the same support and test trials, with PC scores standardized before Ridge adaptation. We computed log-shot relative transfer using Eq. (4), retaining the full Brain response as the reference at every k. Random subsets of the leading 64 Brain PCs provided a control for the benefit of variance-ranked directions. PCA, effective rank, and PC-transfer details are provided in Appendices C.1–C.4.

Correcting Brain-PC coordinates. We tested whether correcting Model responses within the leading $K = 64$ Brain PCs was sufficient to restore transfer. For replacement coordinates $z_i^{*,c}$ under condition c, we constructed

$$M_i^c = M_i + \left[z_i^{*,c} - z_K^{\mathrm{B}}(M_i)\right] U_K^{\mathrm{B}}. \tag{7}$$

This replaces Model coordinates in the Brain subspace while preserving the orthogonal response component.

We compared three replacement conditions. Model+Cal (Model-calibrated coordinate replacement) used an affine Ridge map from full Model responses M_i to source-fMRI coordinates $z_K^{\mathrm{B}}(B_i)$, fitted and validated on paired source training data. The fitted map predicts replacement coordinates from Model responses at support and query images. Oracle Brain-PC coordinate replacement used measured source-fMRI coordinates at both support and query images. Shuffled refitted the coordinate map after permuting training image–coordinate pairs. Each replacement was applied to support and test inputs before refitting the target adaptor.

All coordinate-replacement conditions used the same eight support sizes and NC-r log-shot relative transfer defined above. For condition c, we report the fraction of the Brain–Model transfer gap closed within each target, then average across targets:

$$\mathrm{GC}_t(c) = \frac{b_t^c - b_t^{\mathrm{Model}}}{1 - b_t^{\mathrm{Model}}}, \qquad \mathrm{GC}(c) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{GC}_t(c). \tag{8}$$

Model and full Brain define 0% and 100% closure, respectively. An additional visual-feature coordinate-prediction control is described in Appendix C.8; additional coordinate-recovery readouts are detailed in Appendices C.6–C.7.

PC recovery. PC recovery is the held-out-image correlation between Model and source-fMRI population coordinates in a fixed Brain-PC basis, measuring coordinate fidelity. Figure 4B reports recovery averaged over all 64 Brain PCs, using cross-session Brain-coordinate correlation as the reference. Recovery on the source-transfer-sensitive top 20% of eligible PCs is a supplementary diagnostic; PSS supervises all 64 PCs (Appendices C.5–C.7).

3.4 Population Subspace Supervision

Motivated by the coordinate-replacement analysis, Population Subspace Supervision (PSS) strengthens supervision of the leading Brain-PC coordinates during model training. We use the source-population responses and Brain PCA basis defined above, with $K = 64$. Both the population targets and the PCA basis are constructed exclusively from source training data.

PSS projects the response-prediction error onto the source-population subspace and increases the penalty on errors within it (Appendix Fig. 6). For a predictor with a shared cortical output, the additional loss and total training objective are

$$\mathcal{L}_{\mathrm{PSS}} = \mathbb{E}_i \left[\frac{1}{2K} \sum_{h \in \{L,R\}} \left\| (M_i^h - B_i^h)(U_K^{\mathrm{B},h})^\top \right\|_2^2 \right], \qquad \mathcal{L} = \mathcal{L}_{\mathrm{base}} + \lambda \mathcal{L}_{\mathrm{PSS}}, \tag{9}$$

where h indexes hemispheres, $U_K^{\mathrm{B},h}$ is the Brain-PC basis for hemisphere h, $\mathbb{E}_i$ averages training examples, and $\mathcal{L}_{\mathrm{base}}$ is the trial-level source-response MSE. Here M_i^h denotes the predictor output during training. Model uses $\mathcal{L}_{\mathrm{base}}$, and Model+PSS uses $\mathcal{L}_{\mathrm{base}} + \lambda \mathcal{L}_{\mathrm{PSS}}$. The weight λ controls the strength of the additional supervision. The factor $1/(2K)$ averages over hemispheres and coordinates without variance standardization.

For a shared-output model under balanced source-subject sampling, the base MSE already targets the per-image mean response across source subjects. PSS gives greater weight to prediction errors within its leading PC subspace. The relation to the base objective, extensions to other predictors, and configuration-specific hyperparameters are detailed in Appendix D. Configuration-specific training settings are reported in Appendix D.4.

4 Results

We first evaluate the cross-subject transfer capability of model-predicted fMRI responses using brain-referenced transfer evaluation. We then ask what response-space geometry properties underlie this transfer gap and whether correcting them can restore transfer. Finally, we evaluate Population Subspace Supervision across model configurations and datasets.

4.1 Predicted responses fall short of source-fMRI transfer

We first compared Brain and Model transfer in the primary DINOv2–TBEn configuration (Fig. 2A). Both paths' transfer NC-r improved as the number of target support images increased, but mean Brain transfer remained higher at every support size. Increasing target data therefore improved prediction without eliminating the Brain–Model transfer gap.

We then quantified Model transfer relative to Brain transfer using the relative transfer score (Equation 4, Fig. 2B). Beyond 10 support images, the score decreased with increasing target support and stabilized at approximately 0.80 in the high-shot regime, reflecting a growing relative shortfall as Brain transfer improved. After aggregation across shots, the primary configuration achieved a mean log-shot relative transfer score of 0.835.

We further evaluated whether this gap extended across backbone–predictor configurations. Across all evaluated 12 settings, Model relative transfer ranged from 0.559 to 0.835, with DINOv2–TBEn highest (Fig. 2C). The mean Brain–Model transfer gap also persisted with nonlinear target adaptors in the primary configuration (Appendix E.2). The observed gap therefore extends across the evaluated combinations of visual backbone and response predictor.

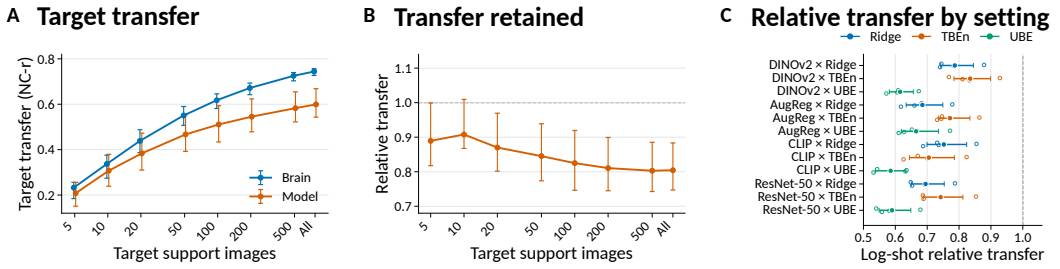


Figure 2: **Baseline model predictions fall short of source-fMRI transfer. A,** DINOv2–TBEn transfer across target support sizes. Transfer NC-r is shown for Brain (blue) and Model (orange). **B,** Model-to-Brain transfer ratio at each support size. In A and B, points are means across four targets, and error bars span the range across the four targets. *All* denotes the maximum available target support. **C,** Log-shot relative transfer across backbone–predictor configurations, colored by predictor. Large points show means across four targets, open points show individual targets, and horizontal intervals show two-sided 95% target-bootstrap confidence intervals. Values below 1 indicate lower transfer than the source-fMRI reference.

4.2 Response geometry and population-coordinate correction

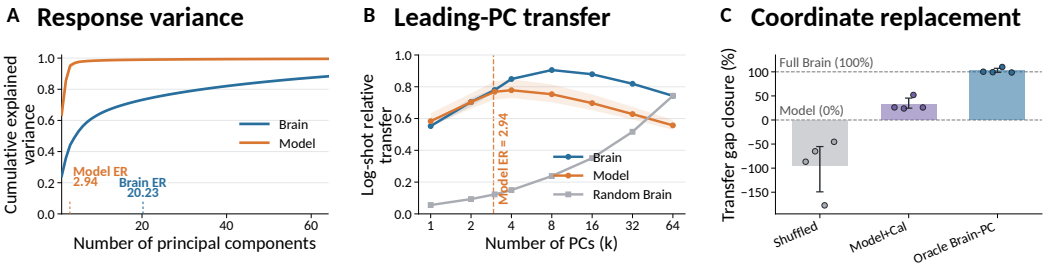


Figure 3: **Population-coordinate correction improves transfer in variance-concentrated Model responses. A,** Cumulative explained variance versus the number of leading PCs on source training images, using each response's own basis and total response variance. Markers indicate entropy-based effective rank within the leading 64 PCs (Brain: 20.23; Model: 2.94). **B,** Log-shot relative transfer $\bar{b}$ versus the number of included PCs: leading Brain PCs, leading Model PCs, and random subsets of the leading 64 Brain PCs. The vertical dashed line marks Model effective rank, and shading shows 95% target-bootstrap intervals. Additional leading Brain PCs support higher transfer than equally sized sets of leading Model PCs. **C,** Brain–Model transfer gap closure after coordinate replacement within the leading 64 Brain PCs, while preserving the orthogonal Model component. Model+Cal predicts Brain-PC coordinates from Model responses; Shuffled refits this map after permuting image–coordinate pairings; Oracle Brain-PC supplies source-fMRI coordinates directly, including at test images, as an oracle intervention. Bars show means across four targets, dots show individual targets, and error bars show 95% target-bootstrap intervals. Dashed lines mark Model (0%) and full Brain (100%); learned correction partly closes the gap, while oracle correction reaches the full-Brain reference.

To understand why Model responses transfer less effectively, we first examined the variance explained by the leading principal components (PCs; Fig. 3A) and found a collapse in the effective rank of the Model responses. The first and first three Brain PCs explained approximately 24% and 44% of the variance, compared with 64% and 95% for Model responses. The effective rank within the leading 64 PCs was 20.23 for Brain and 2.94 for Model. Thus, Model variance was concentrated in far fewer population directions.

We next examined how additional leading PCs contributed to transfer (Fig. 3B). Brain-leading PC inputs supported greater transfer than equally sized Model-leading inputs as more population dimensions were included. Leading Brain PCs also outperformed randomly selected Brain-PC subsets. A complementary spectrum-replacement control showed that imposing the Brain variance spectrum on retained Model directions did not restore transfer (Appendix C.9.1). These results show that additional leading Brain dimensions contain useful structure for cross-subject transfer.

Given that additional leading Brain dimensions carried transfer-relevant structure, we then tested whether correcting the Model's coordinates in this source-population subspace could recover the transfer deficit. We replaced coordinates in the leading 64 Brain PCs while preserving the rest of the Model response (Fig. 3C). Model+Cal replacement predicts Brain coordinates from full Model responses using a map learned on paired source training data; it closed 32.0% of the gap (95% CI $[24.5\%, 45.5\%]$). This improvement shows that source-trained calibration can make existing Model outputs more useful for target adaptation. Oracle Brain-PC replacement supplies source-fMRI coordinates directly, including at test images; it closed 102.1% ($[99.0\%, 107.4\%]$), showing that this oracle correction was sufficient to restore transfer at the brain reference while retaining the orthogonal Model component. Shuffled replacement refits the map after permuting training image–coordinate pairings; its negative closure of -93.6% ($[-149.2\%, -55.2\%]$) shows that correct correspondence matters for the learned correction. A complementary recovery analysis showed that a learned readout improved coordinate prediction over direct Model projection (Appendix C.9.2).

Together, these results associate the transfer deficit with concentrated Model variance and weaker transfer from additional leading Model PCs. Redistributing variance along retained directions did not restore transfer, whereas correcting population coordinates improved it. This pattern motivates stronger supervision of population-coordinate prediction during training.

4.3 Population subspace supervision improves target transfer

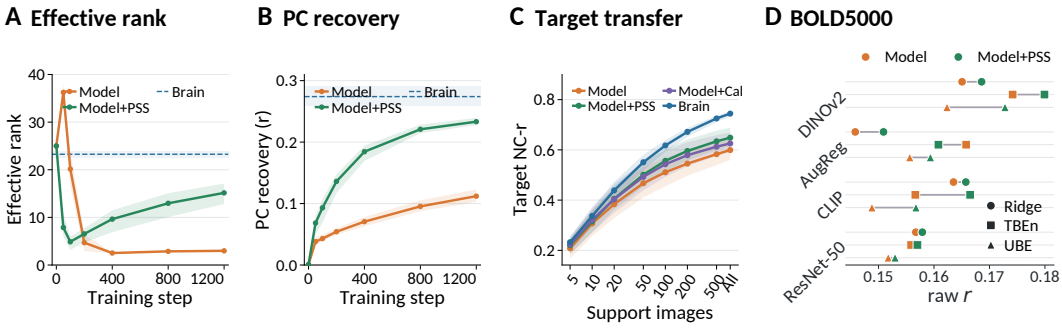


Figure 4: **PSS improves population-coordinate fidelity and target transfer. A,** Effective rank during training for Model (orange) and Model+PSS (green) on held-out images. The dashed blue line shows the Brain effective-rank reference, which is fixed across training steps. **B,** Brain-PC recovery during training for Model and Model+PSS, averaged over all 64 leading Brain PCs. The dashed blue line shows cross-session Brain-coordinate correlation averaged over the same PCs; shading in A and B shows 95% hierarchical target–fold bootstrap intervals. **C,** NSD target-transfer NC-r across support sizes for Model, Model+Cal, Model+PSS, and Brain, shown as four-target means with 95% target-bootstrap bands. *All* denotes the maximum available target support. **D,** Paired BOLD5000 full-support raw r for 12 backbone–predictor configurations. Each line connects Model (orange) and Model+PSS (green) for one configuration; circles, squares, and triangles denote Ridge, TBEn, and UBE, respectively. Raw r is not directly comparable with panel C's NC-r; log-shot AUC results across the full support range are reported in Appendix F.

The coordinate-correction results motivate Population Subspace Supervision (PSS), which increases the penalty on prediction errors within the leading source-population PC subspace to improve transfer (Section 3.4).

We first examined whether PSS improves effective rank (Fig. 4A) and PC recovery (Fig. 4B) on held-out images available in both source-session views (Section 3.3). Model effective rank initially rose from step 0 to 50, then fell sharply and remained low after step 400; Model+PSS effective rank fell through step 100 and then increased steadily. At the final checkpoint (step 1200), effective rank was 2.98 for Model and 15.15 for Model+PSS, compared with 23.25 for the Brain effective rank computed from the mean of the two session views. At the same checkpoint, recovery across all 64 supervised PCs was $r = 0.1120$ for Model and $r = 0.2332$ for Model+PSS, compared with the cross-session Brain-coordinate correlation of $r = 0.2740$ (Appendix C.9.2). Thus, PSS produced

both a less concentrated variance distribution and more faithful population coordinates on these held-out images.

We next examined whether these structural improvements of PSS translated into better target transfer (Fig. 4C). Model+PSS achieved higher mean transfer than Model at all eight support sizes and slightly higher overall transfer than Model+Cal, while remaining below the Brain reference. Mean log-shot relative transfer increased from 0.835 for Model to 0.882 for Model+Cal and 0.899 for Model+PSS. PSS closed 39.5% of the target-level transfer gap on average.

Across all 12 NSD configurations, PSS improved mean full-support NC-r, with varying gains (Table 1), and improved mean log-shot transfer AUC and relative transfer across the support range (Appendix E.3). Its advantage persisted with nonlinear target adaptors in the primary DINOv2–TBEn configuration (Appendix E.2).

To test whether the benefit extends beyond NSD, we further evaluated PSS on BOLD5000 using raw Pearson correlation (Section 3.1). PSS improved mean full-support raw-r in 11 of the 12 configurations (Fig. 4D; Table 1). In the primary DINOv2–TBEn configuration, mean raw-r increased from 0.174 to 0.180. Dataset-specific procedures and support-curve results are provided in Appendix F.

Table 1: **NSD-Shared1000 NC-r and BOLD5000 full-support prediction before and after PSS.** For NSD, Model and Model+PSS give target-mean full-support NC-r. BOLD5000 entries give full-support raw-r mean $\pm$ sample SD across four targets. Both datasets use fixed Ridge adaptor $\alpha = 10$.

Backbone	Predictor	NSD-Shared1000 (NC-r)		BOLD5000 (raw-r)	
		Model	Model+PSS	Model	Model+PSS
DINOv2	Ridge	0.578 ± 0.051	$\mathbf{0.605 \pm 0.048}$	0.165 ± 0.011	$\mathbf{0.169 \pm 0.011}$
	TBEn	0.603 ± 0.052	$\mathbf{0.647 \pm 0.044}$	0.174 ± 0.010	$\mathbf{0.180 \pm 0.010}$
	UBE	0.471 ± 0.059	$\mathbf{0.608 \pm 0.048}$	0.162 ± 0.008	$\mathbf{0.173 \pm 0.007}$
AugReg	Ridge	0.512 ± 0.055	$\mathbf{0.514 \pm 0.057}$	0.146 ± 0.007	$\mathbf{0.151 \pm 0.005}$
	TBEn	0.542 ± 0.049	$\mathbf{0.553 \pm 0.046}$	$\mathbf{0.166 \pm 0.010}$	0.161 ± 0.011
	UBE	0.482 ± 0.057	$\mathbf{0.531 \pm 0.052}$	0.156 ± 0.008	$\mathbf{0.159 \pm 0.009}$
CLIP	Ridge	0.557 ± 0.057	$\mathbf{0.571 \pm 0.056}$	0.164 ± 0.009	$\mathbf{0.166 \pm 0.009}$
	TBEn	0.508 ± 0.064	$\mathbf{0.578 \pm 0.054}$	0.157 ± 0.010	$\mathbf{0.167 \pm 0.007}$
	UBE	0.430 ± 0.057	$\mathbf{0.511 \pm 0.061}$	0.149 ± 0.008	$\mathbf{0.157 \pm 0.007}$
ResNet-50	Ridge	0.516 ± 0.061	$\mathbf{0.526 \pm 0.064}$	0.157 ± 0.008	$\mathbf{0.158 \pm 0.008}$
	TBEn	0.516 ± 0.061	$\mathbf{0.545 \pm 0.055}$	0.156 ± 0.009	$\mathbf{0.157 \pm 0.008}$
	UBE	0.433 ± 0.046	$\mathbf{0.506 \pm 0.051}$	0.152 ± 0.008	$\mathbf{0.153 \pm 0.007}$

Together, PSS improves population-coordinate fidelity and primary-setting target transfer, with prediction gains across the evaluated NSD and most BOLD5000 configurations, providing a practical intervention for reducing the transfer deficit.

5 Conclusion and Limitations

We introduce brain-referenced transfer evaluation and identify a cross-subject transfer gap between model-predicted and source-fMRI responses under matched target adaptation. Geometry analyses reveal concentrated Model variance and greater transfer from leading Brain PCs than equally sized Model-PC sets. Coordinate interventions identify population-coordinate prediction as an improvement target. By upweighting errors in this subspace, Population Subspace Supervision (PSS) improves mean transfer across 12 NSD configurations and mean correlation in 11 of 12 BOLD5000 configurations. These findings support direct transfer evaluation and population-subspace supervision to prioritize transfer alongside source-response prediction during development.

The primary NSD analysis is limited to Shared1000, four participants viewing natural images, and a matched Ridge target adaptor; nonlinear-adaptor and BOLD5000 analyses use separate protocols. The mechanisms behind the transfer gap and PSS gains remain incompletely understood, including the roles of prediction error, regularization, readout flexibility, and model capacity. PSS also leaves a residual Brain–Model gap. Future work should test larger, more diverse source populations, investigate these mechanisms, and refine the supervision objective.

AI USE STATEMENT

We used generative AI tools to assist with research ideation and execution, including feedback on the framing of the research question, experimental design, and interpretation of results. We also used these tools to draft and revise portions of the manuscript, improve the clarity of the writing, and assist with scientific figure design and editing. The authors reviewed and revised AI-assisted content and take full responsibility for the methods, results, claims, and final manuscript.

REFERENCES

Hossein Adeli, Sun Minni, and Nikolaus Kriegeskorte. Transformer brain encoders explain human high-level visual responses. *Advances in neural information processing systems*, 38:56840–56871, 2026.

Jannis Ahlert, Thomas Klein, Felix Wichmann, and Robert Geirhos. How aligned are different alignment metrics? *arXiv preprint arXiv:2407.07530*, 2024.

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, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fmri. *Advances in Neural Information Processing Systems*, 36:21895–21907, 2023.

Roman Beliy, Navve Wasserman, Amit Zalcher, and Michal Irani. The wisdom of a crowd of brains: A universal brain encoder. *arXiv preprint arXiv:2406.12179*, 2024.

Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific data*, 6(1):49, 2019.

Haomiao Chen, Keith Jamison, Mert Sabuncu, and Amy Kuceyeski. Beyond grid-locked voxels: Neural response functions for continuous brain encoding. In *International Conference on Learning Representations*, volume 2026, pp. 142312–142338, 2026.

Po-Hsuan Cameron Chen, Janice Chen, Yaara Yeshurun, Uri Hasson, James Haxby, and Peter J Ramadge. A reduced-dimension fmri shared response model. *Advances in neural information processing systems*, 28, 2015.

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

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829. IEEE, 2023.

SueYeon Chung and Larry F Abbott. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current opinion in neurobiology*, 70:137–144, 2021.

Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature communications*, 15(1):9383, 2024.

Stéphane d'Ascoli, Jérémy Rapin, Yohann Benchetrit, Teon Brooks, Katelyn Begany, Joséphine Raugel, Hubert Banville, and Jean-Rémi King. A foundation model of vision, audition, and language for in-silico neuroscience. *arXiv preprint arXiv:2605.04326*, 2026.

Eric Elmoznino and Michael F Bonner. High-performing neural network models of visual cortex benefit from high latent dimensionality. *PLoS computational biology*, 20(1):e1011792, 2024.

Jenelle Feather, Meenakshi Khosla, N Murty, and Aran Nayebi. Brain-model evaluations need the neuroai turing test. *arXiv preprint arXiv:2502.16238*, 2025.

Umut Güçlü and Marcel AJ 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.

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

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

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.

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*, volume 2026, pp. 112924–112956, 2026.

Eghbal Hosseini and Evelina Fedorenko. Large language models implicitly learn to straighten neural sentence trajectories to construct a predictive representation of natural language. volume 36, pp. 43918–43930, 2023.

Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012.

Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532 (7600):453–458, 2016.

Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008. doi: 10.1038/nature06713. URL https://doi.org/10.1038/nature06713.

Meenakshi Khosla, Gia H Ngo, Keith Jamison, Amy Kuceyeski, and Mert R Sabuncu. A shared neural encoding model for the prediction of subject-specific fmri response. In *International Conference on Medical image Computing and Computer-Assisted Intervention*, pp. 539–548. Springer, 2020.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025.

Nikolaus Kriegeskorte and Rogier A Kievit. Representational geometry: integrating cognition, computation, and the brain. *Trends in cognitive sciences*, 17(8):401–412, 2013.

Nikolaus Kriegeskorte and Xue-Xin Wei. Neural tuning and representational geometry. *Nature reviews neuroscience*, 22(11):703–718, 2021.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Yixuan Liu, Zhiyuan Ma, Likai Tang, Runmin Gan, Xinche Zhang, Jinhao Li, Chao Xie, and Sen Song. Training-driven representational geometry modularization predicts brain alignment in language models. In *Proceedings of the 48th Annual Meeting of the Cognitive Science Society*, 2026. URL https://escholarship.org/uc/item/6c19n50h.

Florian P Mahner, Lukas Muttenthaler, Umut Güçlü, and Martin N Hebart. Dimensions underlying the representational alignment of deep neural networks with humans. *Nature Machine Intelligence*, 7(6):848–859, 2025.

Omer Moussa and Mariya Toneva. Rabbit: Rapidly adaptive bold foundation model via brain-tuning for accurate zero-shot and few-shot prediction of speech-elicited responses in the brain. *arXiv preprint arXiv:2607.05171*, 2026.

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.

Samuel A Nastase, Yun-Fei Liu, Hanna Hillman, Kenneth A Norman, and Uri Hasson. Leveraging shared connectivity to aggregate heterogeneous datasets into a common response space. *NeuroImage*, 217:116865, 2020.

Soma Nonaka, Kei Majima, Shuntaro C Aoki, and Yukiyasu Kamitani. Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *IScience*, 24(9), 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, pp. 407007, 2018.

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

Ansh Soni, Sudhanshu Srivastava, Konrad Kording, and Meenakshi Khosla. Conclusions about neural network to brain alignment are profoundly impacted by the similarity measure. *bioRxiv*, pp. 2024–08, 2024.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

Imran Thobani, Javier Sagastuy-Brena, Aran Nayebi, Jacob Prince, Rosa Cao, and Daniel Yamins. Model-brain comparison using inter-animal transforms. *arXiv preprint arXiv:2510.02523*, 2025.

Navve Wasserman, Roman Beliy, Roy Urbach, and Michal Irani. Functional brain-to-brain transformation with no shared data. *arXiv preprint arXiv:2404.11143*, 2024.

Haiguang Wen, Junxing Shi, Wei Chen, and Zhongming Liu. Transferring and generalizing deep-learning-based neural encoding models across subjects. *NeuroImage*, 176:152–163, 2018a.

Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral cortex*, 28(12): 4136–4160, 2018b.

Jialin Wu, Shreya Saha, Yiqing Bo, and Meenakshi Khosla. Measuring the measures: Discriminative capacity of representational similarity metrics across model families. *arXiv preprint arXiv:2509.04622*, 2025.

Daniel LK 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.

Junjie Yu, Wenxiao Ma, Chen Wei, Jianyu Zhang, Haotian Deng, Zihan Deng, and Quanying Liu. Local intrinsic dimension of representations predicts alignment and generalization in ai models and human brain. *arXiv preprint arXiv:2601.22722*, 2026a.

Muquan Yu, Mu Nan, Hossein Adeli, Jacob Prince, John A Pyles, Leila Wehbe, Maggie Henderson, Michael Tarr, and Andrew Luo. Meta-learning an in-context transformer model of human higher visual cortex. *Advances in Neural Information Processing Systems*, 38:168122–168180, 2026b.

A Detailed experimental protocol

A.1 Data units and fold construction

We used NSD Shared1000 trials from subj01, subj02, subj05, and subj07. The left and right hemispheres contained the common valid fsaverage vertices across subjects and sessions. All operations were performed separately by hemisphere.

We assigned each image i and session q a deterministic five-fold label:

$$f_{\mathrm{img}}(i) = \mathcal{H}(i) \bmod 5, \qquad f_{\mathrm{ses}}(q) = (q-1) \bmod 5, \tag{10}$$

where $\mathcal{H}$ is a fixed hash of the image identifier. For outer fold f, a trial was in the outer-training set only when

$$f_{\mathrm{img}}(i) \neq f \quad \text{and} \quad f_{\mathrm{ses}}(q) \neq f. \tag{11}$$

The primary query contained trials for which both labels equaled f. Therefore, neither the query images nor the query sessions entered predictor training, PCA fitting, support selection, or adaptor fitting. Support images were drawn from the outer-training image pool. The sets were nested within each support order. Repeated trials of a selected support image were included only when they belonged to the outer-training session pool.

Configuration coverage. The 12 evaluated configurations form the full factorial combination of four visual backbones and three predictors.

A.2 Brain and Model source representations

For a held-out target t, let $\mathcal{S}_t$ be the other three subjects. The Brain representation for image i and hemisphere h was

$$B_i^h = \frac{1}{|\mathcal{S}_t|} \sum_{s \in \mathcal{S}_t} \left(\frac{1}{|\mathcal{R}_{s,i}|} \sum_{r \in \mathcal{R}_{s,i}} y_{s,i,r}^h \right), \tag{12}$$

where $\mathcal{R}_{s,i}$ contains the repeats available for image i within the corresponding evaluation split. For training and support images, only outer-training-session repeats are used. At a query image, B_i^h is constructed from the source-subject responses in the held-out query pool.

The Model representation was $M_i^h = f_{\theta_t}(H(x_i))^h$, where x_i is the stimulus image and H is the frozen visual backbone. The predictor was trained on source-subject outer-training responses. Brain and Model occupy the same cortical response space.

Ridge and TBEn produced a shared cortical response. For UBE, we averaged the predictions for the three source subjects before target adaptation. This averaged response served as the Model representation throughout the transfer evaluation.

A.3 Target adaptor and evaluation metric

Let $X_s \in \mathbb{R}^{N_s \times D}$ be a source representation on the target support trials, and let $Y_s \in \mathbb{R}^{N_s \times V}$ be the target fMRI response. We standardized every feature with its support-set mean and standard deviation and divided the result by $\sqrt{D}$. The query representation used the same support statistics. Let the standardized matrices be $\widetilde{X}_s$ and $\widetilde{X}_q$, and let $\bar{Y}_s$ be the support-response mean. The dual Ridge prediction was

$$\widehat{Y}_q = \bar{Y}_s + \widetilde{X}_q \widetilde{X}_s^\top \left(\widetilde{X}_s \widetilde{X}_s^\top + \alpha_{\mathrm{adapt}} I \right)^{-1} (Y_s - \bar{Y}_s), \tag{13}$$

The primal form gives the same prediction when $D < N_s$. All response conditions used matched support and query trials. The support-only tuned analysis selected regularization separately within each support set using Appendix A.4; the displayed fixed-adaptor analyses used $\alpha_{\mathrm{adapt}} = 10$.

We computed vertex-wise Pearson correlation and raw-r using Eq. (1). From the NSD noise-ceiling signal-to-noise ratio c_v, the single-trial correlation ceiling was

$$\mathrm{NC}_v = \frac{c_v}{\sqrt{c_v^2 + 1}}. \tag{14}$$

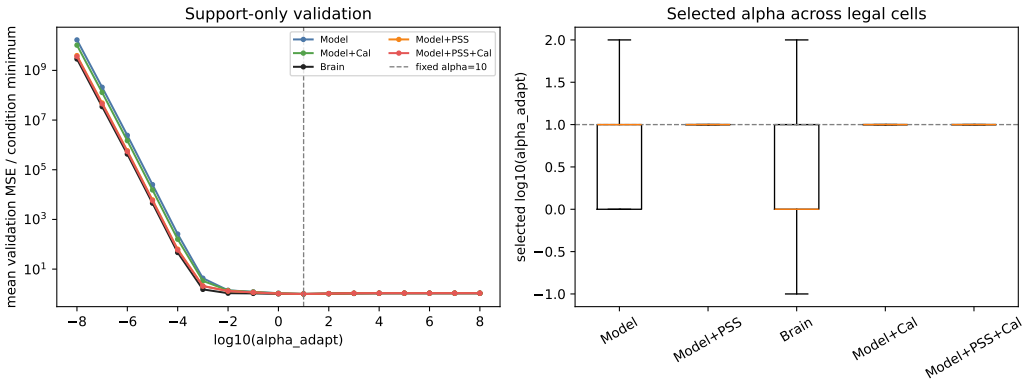


Figure 5: **Support-only Ridge regularization selection.** Left: pooled support-validation MSE relative to each condition's minimum, across the tested α_{adapt} grid. Right: selected α_{adapt} across legal cells. The dashed line marks the fixed value $\alpha = 10$; target query responses were never used for selection.

The main metric was NC-r $= r/(V^{-1}\sum_v \text{NC}_v)$. We first computed this metric within a target, fold, hemisphere, model seed, and support order. Hemispheres and other repeated factors were then averaged within the same target and shot.

A.4 Support-only target-adaptor tuning

The main analyses use a fixed target-adaptor regularization $\alpha_{\text{adapt}} = 10$ to enable matched comparisons across all backbone–predictor configurations. We additionally tested whether the primary DINOv2–TBEn result depends on this fixed choice by selecting α_{adapt} independently for each response condition using only the target support set.

For every target, outer fold, hemisphere, model seed, support order, support size, and input representation, we divided the target support set into at most five deterministic inner folds grouped by image identity, so repeated trials of one image remained in the same fold. We selected α_{adapt} by pooled trial-weighted validation MSE over response vertices. Preprocessing statistics were refitted on each inner-training split, and the outer query was never used for selection. The default grid was $\{10^{-6}, 10^{-5}, \ldots, 10^{6}\}$; when its optimum lay on a boundary, the search admitted two additional decades on that side. The selected adaptor was then refitted on the complete support set.

Under support-only tuning, mean relative transfer increased from 0.8537 for Model to 0.9143 for Model+PSS. The paired NC-r log-shot AUC difference was 0.1589 (95% target-bootstrap CI [0.1012, 0.2145]), with a positive difference for all four targets. PSS closed 43.0% of the target-level transfer gap on average (95% CI [37.1%, 49.5%]).

Appendix Fig. 5 shows the support-validation curves and selected regularization strengths across target, fold, hemisphere, model-seed, support-order, and support-size combinations.

B Transfer aggregation and uncertainty

B.1 Curve aggregation

Let $n_1 < \cdots < n_J$ be the eight support sizes and let $R^c_{t,j}$ be the nuisance-averaged NC-r for target t, shot n_j, and condition $c \in \{\text{Model}, \text{Brain}\}$. The target-level area was computed by the trapezoidal rule on the log-shot axis:

$$A^c_t = \sum_{j=1}^{J-1} \frac{\log n_{j+1} - \log n_j}{2} \left(R^c_{t,j} + R^c_{t,j+1}\right). \tag{15}$$

We then computed $b_t = A_t^{\mathrm{Model}}/A_t^{\mathrm{Brain}}$ and averaged the four target-level ratios to obtain $\bar{b}$. This target-first order keeps each subject equally weighted.

The eight support sizes were 5, 10, 20, 50, 100, 200, 500, and full support. The full-support endpoint was placed at the mean available image count across folds; zero-shot transfer was excluded. On the log-shot axis, an interval contributes width $\log n_{j+1} - \log n_j = \log(n_{j+1}/n_j)$, so equal multiplicative increases in support size receive equal integration weight.

For a condition c, let b_t^c denote its score relative to the full Brain response. The fraction of the target-level Brain–Model transfer gap closed is

$$\mathrm{GC}_t(c) = \frac{b_t^c - b_t^{\mathrm{Model}}}{1 - b_t^{\mathrm{Model}}}. \tag{16}$$

This closure is zero for Model and one for Brain. Spectrum replacement, coordinate replacement, and the PSS transfer-curve analysis report the mean of these target-level closures. Values above one indicate transfer exceeding the Brain reference; negative values indicate transfer below Model. When the Brain area is nonpositive or the Model gap is near zero, the corresponding ratio or closure is flagged rather than clipped.

B.2 Uncertainty and paired comparisons

We first obtain one score or paired difference for each target subject, then sample the four targets with replacement 20,000 times and recompute the target mean. The 2.5th and 97.5th percentiles give the two-sided 95% target-bootstrap interval. All method comparisons resample paired target-level differences; folds, hemispheres, seeds, and support orders are repeated factors within a target rather than independent subjects. The intervals summarize variation across the four evaluated targets, whose LOSO source sets overlap. Plots retain all four target points alongside aggregate intervals.

C Mechanism analysis details

C.1 PCA and response coordinates

PCA was fitted separately for each LOSO target, outer fold, and hemisphere using source outer-training images. For N training images, we computed

$$\mu_{\mathrm{B}} = N^{-1} \sum_{i \in \mathcal{I}_{\mathrm{train}}} B_i, \qquad X_{\mathrm{B}} = [B_i - \mu_{\mathrm{B}}]_{i \in \mathcal{I}_{\mathrm{train}}}, \tag{17}$$

$$\mu_{\mathrm{M}} = N^{-1} \sum_{i \in \mathcal{I}_{\mathrm{train}}} M_i, \qquad X_{\mathrm{M}} = [M_i - \mu_{\mathrm{M}}]_{i \in \mathcal{I}_{\mathrm{train}}}. \tag{18}$$

The leading k right singular vectors of X_{B} and X_{M} form the rows of U_k^{B} and U_k^{M}, respectively. Both bases have orthonormal rows. Coordinates are computed using Eq. (5). In particular, $z_k^{\mathrm{B}}(B_i)$ and $z_k^{\mathrm{M}}(M_i)$ use each response's own basis, and $z_k^{\mathrm{B}}(M_i)$ projects Model responses into the Brain basis. Means and bases remain fixed for support and query images. Coordinate recovery and replacement use $K = 64$. Repeated trials of an image receive the same PC-score input.

C.2 Variance spectra and effective rank

Let s_j^{B} and s_j^{M} be the singular values of X_{B} and X_{M} in descending order. Figure 3A uses cumulative explained variance

$$\mathrm{CEV}_k^{\mathrm{B}} = \frac{\sum_{j=1}^{k} (s_j^{\mathrm{B}})^2}{\|X_{\mathrm{B}}\|_F^2}, \qquad \mathrm{CEV}_k^{\mathrm{M}} = \frac{\sum_{j=1}^{k} (s_j^{\mathrm{M}})^2}{\|X_{\mathrm{M}}\|_F^2}. \tag{19}$$

Effective rank uses the leading-64 variance proportions and entropy definition in Eq. (6). A uniform spectrum over m components has ER m, and concentration in one component has ER one.

For the spectrum-replacement control, these quantities were also recomputed after the support-fitted centering, feature standardization, and $1/\sqrt{D}$ scaling used by the target adaptor. Numerical rank

was bounded by the centered support-matrix structural cap; for example, the five-shot cap was four. This separates the geometry actually presented to Ridge from spectra computed on unstandardized source responses.

C.3 Spectrum replacement

We computed the training singular spectrum of each representation and retained singular values exceeding $\sqrt{\epsilon_{\text{float32}}} = 3.45 \times 10^{-4}$ times the largest singular value. For recipient A and donor D, let q be the smaller numerical rank. The replacement singular values are

$$\widetilde{s}_j^D = s_j^D \frac{\|s_{1:q}^A\|_2}{\|s_{1:q}^D\|_2}, \qquad j = 1, \ldots, q. \tag{20}$$

Writing v_j^A as a column principal direction and $z_{i,j}^A = (a_i^A - \mu_A)v_j^A$, we transform every image using

$$\widetilde{a}_i^A = \mu_A + \sum_{j=1}^{q} z_{i,j}^A \frac{\widetilde{s}_j^D}{s_j^A} (v_j^A)^\top. \tag{21}$$

This preserves the recipient's training mean, retained directions, and variance within the common rank, since $\sum_{j=1}^q (\widetilde{s}_j^D)^2 = \sum_{j=1}^q (s_j^A)^2$. Components beyond q are omitted. The transform is fixed before target adaptation.

We evaluated both directions: Brain spectrum on Model directions and Model spectrum on Brain directions. Each transformed cortical response underwent the same support-fitted standardization and Ridge adaptation as Model and full Brain. For fixed-adaptor transfer evaluation, we averaged NC-r within each target and shot, integrated all eight support sizes over $\log n$, and computed $\mathrm{GC}_t(c)$ from Eq. (16). Points are means of the four target-level closures; intervals use 20,000 bootstrap resamples of targets.

C.4 PC inputs and direction-wise transfer

For $k \in \{1, 2, 3, 4, 8, 16, 32, 64\}$, the three inputs in Figure 3B are

$$q_i^{c,k} = \begin{cases} \left[\dfrac{[z_{64}^{\mathrm{B}}(B_i)]_j - m_j^{\mathrm{B}}}{\max(d_j^{\mathrm{B}}, 10^{-6})} \right]_{j=1}^{k}, & c = \text{Brain-leading}, \\ \left[\dfrac{[z_{64}^{\mathrm{M}}(M_i)]_j - m_j^{\mathrm{M}}}{\max(d_j^{\mathrm{M}}, 10^{-6})} \right]_{j=1}^{k}, & c = \text{Model-leading}, \\ \left[\dfrac{[z_{64}^{\mathrm{B}}(B_i)]_j - m_j^{\mathrm{B}}}{\max(d_j^{\mathrm{B}}, 10^{-6})} \right]_{j \in J_k}, & c = \text{Random Brain}. \end{cases} \tag{22}$$

Here $m_j^{\mathrm{B}}, d_j^{\mathrm{B}}$ and $m_j^{\mathrm{M}}, d_j^{\mathrm{M}}$ are source-training means and standard deviations of the corresponding PC scores, and J_k is a size-k subset of $\{1, \ldots, 64\}$. Random subsets are sampled uniformly without replacement within a subset. We used 100 subsets for each $k < 64$ and the complete set for $k = 64$; each subset was fixed across shots and support orders within a target–fold–hemisphere cell.

Each input was mapped to target trial rows and passed through the same support-set standardization, $1/\sqrt{k}$ scaling, and Ridge adaptor with $\alpha = 10$ described in Appendix A. Adaptors were fitted separately for every representation, subset, and support set. Let $\bar{r}_t^{c,k}(n)$ denote the mean target-test NC-r for condition c with k PC inputs. Its aggregate transfer score is

$$\bar{b}^c(k) = \frac{1}{T} \sum_{t=1}^{T} \frac{\mathrm{AUC}_{\log n}\left(\bar{r}_t^{c,k}(n)\right)}{\mathrm{AUC}_{\log n}\left(\bar{r}_t^{\mathrm{Brain}}(n)\right)}. \tag{23}$$

We averaged NC-r over folds, hemispheres, model seeds, support orders, and random subsets within each target and shot, then applied Eq. (23) over all eight support sizes. The denominator uses the full cortical Brain response, not a PC-restricted Brain input. Figure 3B plots $\bar{b}^c(k)$, with 95% intervals from 20,000 target-bootstrap resamples. The plotted names Brain, Model, and Random correspond to Brain-leading, Model-leading, and Random Brain, respectively.

C.5 Selection of transfer-relevant Brain PCs

As a supplementary diagnostic, we selected a source-transfer-sensitive subset $\mathcal{J}$ of the leading 64 Brain PCs. For each PC, eligibility required mean source-to-population coordinate correlation above 0.5, even–odd source-session coordinate correlation above 0.1, and outer-training-only DINOv2-predicted coordinate correlation above 0.1. Visual predictability used four held-in image folds for out-of-fold prediction, with Ridge regularization selected by a nested rotation over the remaining three folds. The outer query was not loaded during selection.

To rank eligible PCs by source-transfer sensitivity, we predicted each source subject's full cortical response from the other sources' mean Brain-PC scores. If Z_{-s} contains these scores and Y_s the responses of source s, with both matrices centered, the map and sensitivity matrix are

$$W_s = (Z_{-s}^\top Z_{-s} + \alpha_G I)^{-1} Z_{-s}^\top Y_s, \qquad G_{\text{transfer}} = \frac{1}{|\mathcal{S}_t|} \sum_{s \in \mathcal{S}_t} W_s W_s^\top, \tag{24}$$

where $\alpha_G = 1000$. The diagonal scores rank input PCs by the squared magnitude of their fitted source-response readouts. Among N_{elig} eligible PCs, ascending rank j receives percentile $(j - \frac{1}{2})/N_{\text{elig}}$; $\mathcal{J}$ contains those with percentile at least 0.8. The same selected PC indices are used for every recovery condition and model seed within a target–fold–hemisphere cell. They differ from the variance-ranked sets in Appendix C.4; coordinate replacement uses all 64 PCs. Across the 40 target–fold–hemisphere cells, 39–48 PCs were eligible and 8–10 were selected, with a mean of 9.05 selected PCs per cell.

Recovery on source-selected PCs. At training step 1200, Model and Model+PSS achieved $R_{\text{PC},\mathcal{J}} = 0.1002$ and 0.2513, respectively. The cross-session Brain reference was 0.2750. Source-trained coordinate readouts from DINOv2 features and Model predictions achieved 0.2971 and 0.3053. Figure 4B instead reports recovery across all 64 PCs; the corresponding values are detailed in Appendix C.9.2.

C.6 Source-trained coordinate maps

Let B_i^ℓ denote the even-session, odd-session, or full-population Brain response, with $B_i^{\text{pop}} = B_i$. Its coordinates in the fixed Brain basis are

$$z_i^\ell = (B_i^\ell - \mu_{\text{B}})(U_{64}^{\text{B}})^\top. \tag{25}$$

For input representation A, we fitted an affine Ridge map using source training images $\mathcal{I}_{\text{src}}$:

$$\begin{aligned} \mathcal{R}_{A\to\ell}(A_i) &= \widetilde{A}_i W_{A\to\ell} + d_{A\to\ell}, \\ (W_{A\to\ell}, d_{A\to\ell}) &= \arg\min_{W,d} \left[\sum_{i \in \mathcal{I}_{\text{src}}} \|\widetilde{A}_i W + d - z_i^\ell\|_2^2 + \alpha \|W\|_F^2 \right]. \end{aligned} \tag{26}$$

Here $\widetilde{A}_i$ denotes the standardized and dimension-scaled input, with preprocessing and regularization selected as described below.

For opposite session views a, b, the four recovery estimates were

$$\widehat{z}_i^{c,b} = \begin{cases} (M_i - \mu_{\text{B}})(U_{64}^{\text{B}})^\top, & c = \text{direct}, \\ \mathcal{R}_{M\to b}(M_i), & c = \text{M}, \\ \mathcal{R}_{B^a\to b}(B_i^a), & c = \text{B}, \\ \mathcal{R}_{H\to b}(H(x_i)), & c = \text{H}. \end{cases} \tag{27}$$

They correspond to direct Model projection, learned Model readout, cross-session Brain readout, and readout from frozen DINOv2 features. All four were evaluated against the same coordinates z_i^b.

Equation (26) fits an affine map from source inputs to source-fMRI coordinates. If D_A is the retained input dimension, $W_{A\to\ell} \in \mathbb{R}^{D_A \times 64}$ and $d_{A\to\ell} \in \mathbb{R}^{64}$ are the readout weights and intercept. For recovery, $\ell = b$ is a destination session view and the inputs are $A \in \{M, B^a, H\}$, with a the opposite session view. For replacement, $\ell = \text{pop}$ and $A \in \{M, H\}$. For the visual-feature condition, $A_i = H(x_i)$. Each input–destination pairing has its own fitted map and regularization strength.

The direct projection in Eq. (27) has no learned readout: it uses the fixed Brain mean and PCA basis. The three learned estimates use exactly the input indicated in their map: the full Model response M_i, source-fMRI response B_i^a, or frozen DINOv2 features $H(x_i)$. All recovery estimates are evaluated against the same $z_i^b = (B_i^b - \mu_{\mathrm{B}})(U_{64}^{\mathrm{B}})^\top$, and we repeat the comparison for even-to-odd and odd-to-even views. An additional direct cross-session control projects B_i^a onto the Brain basis without fitting a map.

We restricted comparisons to common image IDs available in both session views for all three source subjects. Training responses exclude the outer held-out session fold. Query images belong to the held-out image fold and use source-fMRI responses grouped by session parity. All conditions share the same training/query image IDs, fixed Brain basis, and coordinate targets.

For an outer image fold $f \in \{0, \ldots, 4\}$, the inner validation fold is $(f+1) \bmod 5$ within the source training pool; the other training images fit the candidate maps. Input features with training standard deviation at most 10^{-6} are removed, and retained features are standardized and divided by the square root of their count. These statistics are fitted separately on the inner fit set for selection and the full source training set for the final map. For standardized inputs A, coordinate targets Z, and test inputs A_q, Ridge predicts

$$\widehat{Z}_q = \bar{Z} + A_q A^\top (AA^\top + \alpha I)^{-1}(Z - \bar{Z}). \tag{28}$$

We selected α from $\{10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}, 3\times10^{-2}, 10^{-1}, 0.3, 1, 3, 10, 30, 100\}$ by the mean Fisher-transformed validation correlation across all 64 output PCs, then refitted on all source training images. Maps were fitted separately for each input and destination session view. Neither map fitting nor regularization selection used held-out-image responses or target-subject fMRI.

The Model and Model+PSS calibration analyses used source-only validation and the declared in-sample encoder-prediction protocol. Recovery uses the fixed Brain basis; Figure 4B uses all 64 Brain PCs, while the source-selected subset in Appendix C.5 is a supplementary diagnostic.

C.7 Recovery metric and uncertainty

For one session direction, hemisphere, and model seed, all-64 recovery is

$$R_{\mathrm{PC}}^c = \frac{1}{64}\sum_{j=1}^{64} \mathrm{corr}_{i\in\mathcal{I}_{\mathrm{query}}}(\widehat{z}_{i,j}^{c,b}, z_{i,j}^b). \tag{29}$$

For the supplementary source-selected subset, we compute

$$R_{\mathrm{PC},\mathcal{J}}^c = \frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}} \mathrm{corr}_{i\in\mathcal{I}_{\mathrm{query}}}(\widehat{z}_{i,j}^{c,b}, z_{i,j}^b). \tag{30}$$

PC correlations are averaged within each session direction, hemisphere, and model seed. These factors then receive equal weight within each target–fold cell, and cell averages receive equal weight in the reported mean. Figure 4B uses the all-64 metric; the source-selected subset is reported as a supplementary diagnostic in Appendices C.5 and C.9.2. Confidence intervals use 5,000 hierarchical bootstrap samples, resampling targets and then folds within each sampled target. The Brain reference uses direct cross-session coordinates with the same PCs, images, and aggregation.

C.8 Subspace coordinate replacement and gap closure

Coordinate replacement used all 64 leading Brain PCs. The replacement coordinates were

$$z_i^{*,c} = \begin{cases} \mathcal{R}_{M\to\mathrm{pop}}(M_i), & c = \text{Model+Cal}, \\ \mathcal{R}_{H\to\mathrm{pop}}(H(x_i)), & c = \text{DINOv2-calibrated}, \\ z_i^{\mathrm{pop}}, & c = \text{Oracle Brain-PC}, \\ \mathcal{R}^{\pi}_{M\to\mathrm{pop}}(M_i), & c = \text{Shuffled}. \end{cases} \tag{31}$$

Here π permutes the image-to-coordinate assignments in the source training set. The four conditions correspond to *Model+Cal*, *DINOv2-calibrated*, *Oracle Brain-PC*, and *Shuffled* in the reported results.

Equation (7) uses the complete $U_{64}^{\mathrm{B}} \in \mathbb{R}^{64\times V}$, not only $\mathcal{J}$. With $\Pi_{\mathrm{B}} = (U_{64}^{\mathrm{B}})^{\top} U_{64}^{\mathrm{B}}$, orthonormality gives

$$(M_i^c - \mu_{\mathrm{B}})(U_{64}^{\mathrm{B}})^{\top} = z_i^{*,c}, \qquad M_i^c(I - \Pi_{\mathrm{B}}) = M_i(I - \Pi_{\mathrm{B}}). \tag{32}$$

Thus the intervention sets exactly the selected subspace coordinates and preserves the orthogonal response component.

For coordinate replacement, we fit $\mathcal{R}_{M\to\mathrm{pop}}$ and $\mathcal{R}_{H\to\mathrm{pop}}$ using the standardization, source-only validation, and refitting procedure in Appendix C.6. Their targets are full-population coordinates $z_i^{\mathrm{pop}} = (B_i - \mu_{\mathrm{B}})(U_{64}^{\mathrm{B}})^{\top}$, rather than coordinates from one session view. Thus the two maps predict the same Brain coordinates from Model responses and visual features, respectively.

The shuffled map $\mathcal{R}^{\pi}_{M\to\mathrm{pop}}$ uses the same Model inputs and selected regularization as the unshuffled Model map, but replaces each training target by $z^{\mathrm{pop}}_{\pi(i)}$. It refits both weights and intercept after permuting image-to-coordinate assignments; the permutation is fixed within each target–fold–hemisphere–model-seed cell. The fourth intervention uses z_i^{pop} directly from source-fMRI responses, including at query images, as an oracle diagnostic of coordinate sufficiency. All replacements in Eq. (31) are applied to both support and query inputs before fitting separate target adaptors.

Coordinate replacement and PSS transfer were evaluated at $n \in \{5, 10, 20, 50, 100, 200, 500, \mathrm{full}\}$ using fixed $\alpha = 10$ and the same NC-r log-shot relative transfer as spectrum replacement and PC transfer. Within each target and support size, we first averaged over five folds, three model seeds, five support orders, and both hemispheres. We placed the full-support endpoint at the mean available image count across folds and integrated the resulting eight-point curve by the trapezoidal rule on $\log n$. Using the target-level quantities in Eq. (16), we averaged target-level gap closure as defined in Eq. (8). All conditions share the same Brain reference and paired support and query samples. Figure 3C and the curve-level summary for Figure 4C use this definition.

For transfer curves, paired differences, and gap closure, we used 20,000 paired bootstrap resamples of the four target subjects, retaining all conditions and support sizes together within each sampled target. Each resample averaged the selected target-level statistics; the 2.5th and 97.5th percentiles define the 95% confidence interval. This is the same bootstrap unit used for spectrum replacement and PC transfer.

C.9 Supplementary mechanism results

C.9.1 Redistributing variance along existing directions

Spectrum replacement redistributes variance within retained response directions using Appendix C.3. In this fixed-adaptor sensitivity analysis, common-rank truncation changed NC-r log-shot AUC by 0.0007 for Model (95% CI $[-0.0007, 0.0021]$) and 0.0007 for Brain ($[-0.0043, 0.0102]$), both negligible at the reported precision. Replacing the Model spectrum with the Brain spectrum reduced AUC by 0.0329 ($[-0.0436, -0.0122]$), while replacing the Brain spectrum with the Model spectrum reduced AUC by 0.0996 ($[-0.1171, -0.0711]$), each relative to the corresponding original recipient. Imposing the Brain spectrum on Model directions did not restore transfer, while imposing the Model spectrum on Brain directions reduced transfer.

C.9.2 Independent recovery of transfer-relevant Brain coordinates

At training step 1200, Figure 4B's direct projection onto all 64 Brain PCs yielded $R_{\mathrm{PC}} = 0.1120$ for Model and 0.2332 for Model+PSS, with a cross-session Brain reference of 0.2740. On the supplementary source-selected subset, the corresponding $R_{\mathrm{PC},\mathcal{J}}$ values were 0.1002 and 0.2513, with a Brain reference of 0.2750. Source-trained coordinate readouts from DINOv2 features and Model responses achieved subset recovery of 0.2971 and 0.3053, respectively. Source-only PC selection and the recovery metric are defined in Appendices C.5–C.7.

C.9.3 Complete coordinate-replacement results

In the fixed-adaptor analysis, all four replacement conditions modify the leading 64 Brain PCs while preserving the orthogonal Model response (Appendix C.8). Predicting replacement coordinates directly from DINOv2 features closed 19.9% of the transfer gap (95% CI $[4.6\%, 34.1\%]$), compared with 32.0% ($[24.5\%, 45.5\%]$) when predicting them from Model responses. This comparison

tests the input used for the source-trained coordinate readout. Shuffled pairing yielded -93.6% ($[-149.2\%, -55.2\%]$), while Oracle Brain-PC replacement yielded 102.1% ($[99.0\%, 107.4\%]$). A closure above 100% indicates a point estimate exceeding the Brain reference; its interval here includes 100%. Oracle Brain-PC is an oracle replacement using source-fMRI coordinates at support and test images, with the Model's orthogonal component retained. It is therefore a hybrid response, and its performance is not the transfer performance of the Brain-PC scores alone.

D Population Subspace Supervision: Details and Derivation

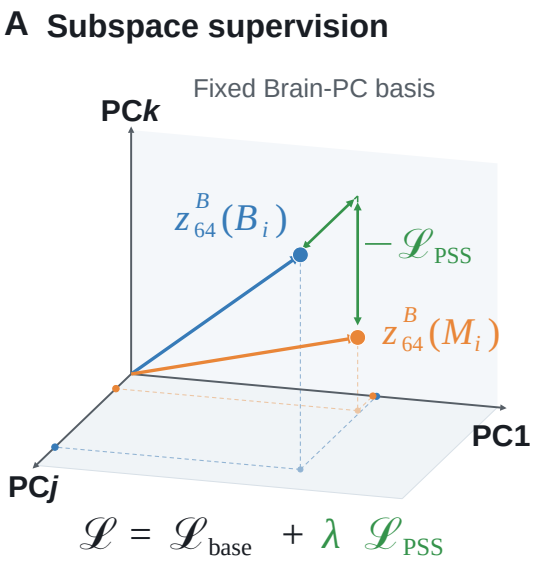


Figure 6: **Population Subspace Supervision in a fixed Brain-PC basis.** Source-fMRI responses B_i and model predictions M_i are projected onto the same leading 64 source-population PCs. The green segments illustrate coordinate discrepancies penalized by $\mathcal{L}_{\text{PSS}}$. This squared coordinate-error penalty is added to the base loss with weight λ (Eq. (9)). Both the Brain targets and the fixed PCA basis are constructed from source training data.

D.1 Population Teacher Construction

For each outer fold and hemisphere, we first averaged repeated responses within each source subject and then averaged across the three source subjects. We fitted a centered rank-K PCA basis to these source-population means, with $K = 64$. The PC scores were not variance-standardized. Both the teacher and the PCA basis used source outer-training data only.

Let $U_h = U_K^{\mathrm{B},h} \in \mathbb{R}^{K \times V_h}$ contain the orthonormal Brain PCA rows for hemisphere h, and let μ_h be the centering mean across source training images of their source-population responses. The score map is $z_h(a) = (a - \mu_h)U_h^\top$. Since the same μ_h is used for the Model prediction M_i^h and Brain teacher B_i^h,

$$z_h(M_i^h) - z_h(B_i^h) = (M_i^h - B_i^h)U_h^\top. \tag{33}$$

PSS increases supervision of errors within the fixed source-population subspace.

Subspace size and PC-transfer analysis. We used $K = 64$ for coordinate replacement and PSS supervision. This shared subspace covers population directions beyond the dominant Model PCs. The leading-k transfer analysis evaluates target adaptation from k PC coordinates, coordinate replacement modifies these coordinates within the cortical response, and PSS supervises their prediction during training. Figure 4B reports all-64 recovery; Appendix C.5 reports the supplementary source-selected analysis, with both summarized in Appendix C.9.2.

D.2 Relation to the base MSE

Consider one hemisphere and omit h for clarity, writing $U = U_K^{\mathrm{B}}$. Let S be the number of source subjects, V the number of vertices, M_i the shared Model response, and $B_i = S^{-1}\sum_s y_{s,i}$ the mean Brain response. The source-averaged base MSE satisfies

$$\frac{1}{SV}\sum_{s=1}^{S}\|M_i - y_{s,i}\|_2^2 = \frac{1}{V}\|M_i - B_i\|_2^2 + C_i, \tag{34}$$

where $C_i = (SV)^{-1}\sum_s \|y_{s,i} - B_i\|_2^2$ does not depend on the model. Thus, under balanced shared-image training, the base loss already has the population mean as its pointwise optimum. PSS changes the relative weights assigned to prediction errors across response directions.

Let $e_i = M_i - B_i$, $\Pi = U^\top U$, $e_i^{\parallel} = e_i\Pi$, and $e_i^{\perp} = e_i(I - \Pi)$. Combining the base term with the hemisphere contribution to PSS gives, up to a model-independent constant,

$$\mathcal{L}_i = \frac{1}{V}\|e_i\|_2^2 + \frac{\lambda}{2K}\|e_i U^\top\|_2^2 + C_i \tag{35}$$

$$= \frac{1}{V}\|e_i^{\perp}\|_2^2 + \left(\frac{1}{V} + \frac{\lambda}{2K}\right)\|e_i^{\parallel}\|_2^2 + C_i. \tag{36}$$

Errors inside the population subspace therefore receive the relative weight

$$1 + \rho, \qquad \rho = \frac{\lambda V}{2K}, \tag{37}$$

while errors outside the subspace keep their base weight.

In a per-source view, let $z_{s,i} = z(y_{s,i})$ and $\bar{z}_i = z(B_i)$. The coordinate-level objective can be written as

$$\|\widehat{z}_i - z_{s,i}\|_2^2 + \rho\|\widehat{z}_i - \bar{z}_i\|_2^2 = (1+\rho)\|\widehat{z}_i - \widetilde{z}_{s,i}\|_2^2 + C, \tag{38}$$

with $\widetilde{z}_{s,i} = (z_{s,i} + \rho\bar{z}_i)/(1+\rho)$. The effective target is therefore pulled toward the source-population mean.

In the balanced shared-output setting, averaging the leave-one-source-out teacher losses gives the same gradient for the auxiliary term because the mean teacher response is B_i.

D.3 Population-coordinate supervision and effective rank

The effect is exact for a linear predictor with Ridge regularization. For one supervised PC, let X be the visual design matrix, w its readout, z an individual source target, and $\bar{z}$ the population target. The objective is

$$\|Xw - z\|_2^2 + \rho\|Xw - \bar{z}\|_2^2 + \alpha\|w\|_2^2. \tag{39}$$

After completing the square, its solution is

$$w_{\mathrm{PSS}} = \left(X^\top X + \frac{\alpha}{1+\rho}I\right)^{-1} X^\top \widetilde{z}. \tag{40}$$

PSS therefore lowers the effective regularization inside the population subspace and uses a less subject-specific target. If σ_d is a singular value of X, the signal shrinkage factor changes from

$$\frac{\sigma_d^2}{\sigma_d^2 + \alpha} \quad \text{to} \quad \frac{\sigma_d^2}{\sigma_d^2 + \alpha/(1+\rho)}. \tag{41}$$

The second factor is no smaller. Stable population coordinates are therefore less likely to be suppressed by regularization. In a nonlinear predictor, the same weighting appears as a larger gradient along the retained Brain subspace.

Recovering image-dependent signals along additional population directions can spread response variance and increase ER. We evaluate this change using ER and held-out Brain-PC recovery.

The primary DINOv2–TBEn transfer curves use step-1200 predictions from three training seeds, eight support sizes, and fixed target-adaptor $\alpha = 10$. The separately reported tuned aggregates reuse these predictions and select target-adaptor regularization within support (Appendix A.4).

Checkpoint analyses evaluated steps 0, 50, 100, 200, 400, 800, and 1200 on held-out images available in both source-session views. Figure 3A uses source-training images, and Figures 4A–B use held-out images. The Brain ER reference in Figure 4A uses the mean of the two source-session views. Figure 4B uses all 64 Brain PCs and direct cross-session coordinate correlation as its Brain reference. ER is computed for each response matrix before aggregation. Recovery correlations are averaged across all 64 PCs before equal-weight aggregation over hemispheres, model seeds, and session views within each target–fold cell. Confidence intervals use 5,000 hierarchical target–fold bootstrap samples. Target-transfer aggregation and uncertainty follow Appendix C.8.

D.4 PSS TRAINING SETTINGS

The PSS weight λ controls the strength of the auxiliary population-subspace supervision and is distinct from the target-adaptor regularization α_{adapt}.

Primary configuration. The primary Shared1000 DINOv2–TBEn configuration uses $K = 64$, $\lambda = 0.25$, no warmup, and 1,200 training updates. The target-adaptor analyses reuse these trained checkpoints.

Configuration-specific settings. For the broader NSD comparison, configurations use $\lambda = 0.25$ without warmup except for AugReg–Ridge and ResNet-50–Ridge ($\lambda = 0.5$), AugReg–TBEn ($\lambda = 0.05$ with 600-step linear warmup), and AugReg–UBE ($\lambda = 0.05$ with 450-step linear warmup). Ridge predictors use a closed-form solution and therefore do not use a training warmup.

BOLD5000 settings. All 12 BOLD5000 configurations use $K = 64$, $\lambda = 0.25$, and fixed target-adaptor regularization $\alpha_{\mathrm{adapt}} = 10$.

E ADAPTATION AND PRACTICAL-BASELINE CONTROLS

E.1 DIRECT VISUAL BASELINES

We tested two alternatives under the same linear target-adaptation budget. Direct frozen DINOv2 features reached mean relative transfer 0.7424, with NC-r log-shot AUC 0.2988 below Model. Predicting the leading 64 Brain-PC coordinates from DINOv2 features before target adaptation reached 0.6667, with AUC 0.5007 below Model. Both alternatives were below Model for each of the four targets. Model responses supported higher transfer than both direct visual features and visual-feature predictions of Brain-PC coordinates.

E.2 SHARED1000 ENCODER / PRIMARY-CHECKPOINT ADAPTOR ROBUSTNESS

Table 2: **Shared1000 primary-checkpoint cross-adaptor validation.** Brain, Model, and Model+PSS columns are target-mean NC-r log-shot AUC. Each contrast shows the paired mean and its 95% target-bootstrap interval; its target-level sign was positive in 4/4 targets for every row. This does not imply a positive difference at every shot.

Adaptor	Brain	Model	Model+PSS	Brain − Model	Brain − (Model + PSS)	Model + PSS − Model
Ridge	2.6679	2.2865	2.4438	$+0.3814\ [0.2157, 0.5024]$	$+0.2241\ [0.1160, 0.2942]$	$+0.1574\ [0.0997, 0.2118]$
RBF-KRR	2.7000	2.2854	2.4371	$+0.4146\ [0.2640, 0.5141]$	$+0.2628\ [0.1655, 0.3156]$	$+0.1517\ [0.0984, 0.1996]$
Residual MLP	2.6424	2.2641	2.4147	$+0.3786\ [0.2157, 0.4998]$	$+0.2280\ [0.1219, 0.2919]$	$+0.1506\ [0.0933, 0.2079]$

We reused the primary DINOv2–TBEn Model and Model+PSS step-1200 encoders without retraining: both were trained only on Shared1000, and Model+PSS used the original source-PC64 objective with $\lambda = 0.25$. The Brain input was the same leave-target-out source-fMRI population reference. We evaluated four targets, five image/session outer folds, three encoder training seeds for Model and Model+PSS, eight support sizes (5, 10, 20, 50, 100, 200, 500, and full), and five fixed support orders. Brain was evaluated once per target and fold, not replicated over encoder seeds. All representations and adaptors used paired target support and the primary query that jointly held out image and session.

Ridge, exact Gaussian RBF kernel Ridge, and a source-PC64 residual MLP were evaluated under the same target-support protocol. For RBF-KRR, the kernel width was 0.25, 1, or 4 times the median positive squared distance among inner-fit images, with $\alpha \in \{10^{-5}, 10^{-3}, 10^{-1}, 1, 10, 100\}$. The MLP used a 64-dimensional residual with hidden width 16 or 32, a zero-initialized output layer, and support-selected weight decay and stopping epoch. Hyperparameters were selected only within target support by image-grouped cross-validation; preprocessing was refitted on each inner-training split and again on the full support set before query prediction.

Brain exceeded Model in target-mean log-shot AUC under all three adaptors, and PSS improved Model transfer under each adaptor. Individual low-shot exceptions occurred for subj05 at 5 and 10 shots with Ridge and residual MLP, and at 10 shots with RBF-KRR.

E.3 Complete Shared1000 transfer comparison

We additionally summarize transfer across the full support range for all 12 NSD Shared1000 backbone–predictor configurations. For each condition, target-transfer NC-r was averaged within target and support size and then integrated over the eight support sizes on the log-shot axis, following Appendix B. Relative transfer $\bar{b}$ uses the corresponding full-Brain transfer AUC as the reference, as defined in Eq. (4).

Across all 12 backbone–predictor configurations, Model+PSS achieved higher mean log-shot transfer AUC than Model. The corresponding log-shot relative transfer score also increased in all 12 configurations. Model+PSS improved transfer for all four targets in 11 of the 12 configurations, with AugReg–TBEn improving for three of four targets. These results show that the PSS benefit is not restricted to the full-support endpoint reported in Table 1, but extends across the target-support range.

F External validation on BOLD5000

We evaluated four BOLD5000 Release 2.0 participants (CSI1–CSI4) using four-subject leave-one-out transfer and five deterministic image/session outer folds. The all-subject image intersection contained 2,952 images. Source responses were registered to a common fsaverage surface, standardized within session, and restricted to 17,512 left-hemisphere and 18,816 right-hemisphere common-valid vertices. A frozen fsaverage anatomical query partition replaced NSD-specific ROI files.

For each target, the model was trained only on the other three participants' legal outer-training data. Target support selected unique images and included all repeats belonging to the legal outer-training session pool. The eight support sizes were 5, 10, 20, 50, 100, 200, 500, and all legal support images. The primary query held out both image and session folds.

The primary external endpoint is full-support raw-r, computed as the mean vertex-wise Pearson correlation in Eq. (1). BOLD5000 does not provide noise-ceiling estimates comparable to the NSD measure used here, and its limited repeat structure makes reliable vertex-wise estimation difficult; we therefore report raw Pearson correlation. We average repeated factors within each target and report the mean and sample SD across the four targets. Figure 4D and Table 1 use this endpoint. Supplementary support-curve analyses summarize raw-r over the log-shot axis.

All 12 BOLD5000 configurations use $K = 64$, PSS weight $\lambda = 0.25$, and target-adaptor regularization $\alpha = 10$. PSS improves mean full-support raw-r in 11 of the 12 configurations. Table 4 provides supplementary log-shot AUC results across the full support range.

In the DINOv2–TBEn configuration, Model and Model+PSS used paired initialization and training streams across four targets, five folds, and three model seeds. Both used Ridge adaptors with $\alpha = 10$ and five support orders. Mean full-support raw-r increased from 0.1742 to 0.1799, a paired gain of 0.0057. Supplementary raw-r log-shot AUC increased from 0.7212 to 0.7373.

Table 3: **Complete NSD Shared1000 transfer comparison.** Entries report target-mean NC-r log-shot AUC and log-shot relative transfer $\bar{b}$ for Model and Model+PSS across the 12 evaluated backbone–predictor configurations. Δ values are Model+PSS minus Model. Wins count targets with a positive paired AUC difference.

Backbone	Predictor	Model AUC	Model+PSS AUC	Δ AUC	Wins	Model $\bar{b}$	Model+PSS $\bar{b}$	$\Delta\bar{b}$
DINOv2	Ridge	2.151	2.344	+0.193	4/4	0.787	0.857	+0.071
DINOv2	TBEn	2.284	2.456	+0.172	4/4	0.835	0.899	+0.064
DINOv2	UBE	1.687	2.332	+0.644	4/4	0.617	0.853	+0.236
AugReg	Ridge	1.872	2.054	+0.183	4/4	0.685	0.752	+0.067
AugReg	TBEn	2.108	2.154	+0.046	3/4	0.771	0.788	+0.017
AugReg	UBE	1.805	2.068	+0.263	4/4	0.660	0.756	+0.097
CLIP	Ridge	2.057	2.244	+0.187	4/4	0.752	0.821	+0.069
CLIP	TBEn	1.930	2.238	+0.308	4/4	0.705	0.819	+0.113
CLIP	UBE	1.530	1.976	+0.446	4/4	0.559	0.722	+0.163
ResNet-50	Ridge	1.902	2.101	+0.199	4/4	0.695	0.768	+0.073
ResNet-50	TBEn	2.018	2.103	+0.085	4/4	0.737	0.769	+0.032
ResNet-50	UBE	1.623	1.969	+0.345	4/4	0.593	0.720	+0.127

Table 4: **Complete BOLD5000 backbone–predictor comparison.** The headline columns are target-mean raw-r log-shot AUC and the paired $\mathrm{Model} + \mathrm{PSS} - \mathrm{Model}$ difference. Wins count targets with a positive AUC difference. The full-support raw-r difference and relative transfer score difference $\Delta\bar{b}$ are secondary diagnostics; the relative transfer score uses a noisy cross-subject fMRI denominator and is not interpreted as a ceiling-relative score.

Backbone	Predictor	Model AUC	Model+PSS AUC	Δ AUC	Wins	Δ full raw r	$\Delta\bar{b}$
DINOv2	Ridge	0.6633	0.6939	+0.0306	4/4	+0.0035	+0.0542
DINOv2	TBEn	0.7212	0.7373	+0.0162	4/4	+0.0057	+0.0290
DINOv2	UBE	0.6752	0.7123	+0.0372	4/4	+0.0105	+0.0665
AugReg	Ridge	0.5641	0.6025	+0.0384	4/4	+0.0051	+0.0686
AugReg	TBEn	0.6828	0.6610	−0.0218	0/4	−0.0050	−0.0394
AugReg	UBE	0.6455	0.6534	+0.0078	2/4	+0.0037	+0.0133
CLIP	Ridge	0.6455	0.6779	+0.0324	4/4	+0.0022	+0.0581
CLIP	TBEn	0.6503	0.6867	+0.0364	4/4	+0.0099	+0.0661
CLIP	UBE	0.6161	0.6422	+0.0260	4/4	+0.0079	+0.0460
ResNet-50	Ridge	0.6087	0.6475	+0.0388	4/4	+0.0012	+0.0691
ResNet-50	TBEn	0.6549	0.6304	−0.0246	0/4	+0.0011	−0.0435
ResNet-50	UBE	0.6269	0.6285	+0.0016	3/4	+0.0012	+0.0024