RELATIVE JACOBIAN NONLINEARITY FOR INTERPRETING NEURAL ENCODING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Similar neural responses can accompany different local dependencies on visual features. Describing these dependencies requires a measure of how an encoding model's mapping changes across images. We introduce *relative Jacobian nonlinearity* (RJN), which compares two local mappings on the scale of their combined sensitivity before averaging over images. This pairwise construction connects overall variation to magnitude disparity, weighted directional disagreement, and the stimuli contributing to each. In visual encoding models, the decomposition locates a fusiform contrast in directional variation, while stimulus scores distinguish mapping profiles among images with closely matched responses. Pair-relative and energy-based scores give almost the same voxel ordering but different image rankings; total RJN profiles are more consistent across alternative fits in the evaluated models. Regional contrasts vary with output reliability and sensitivity, and changing the feature perturbations can reverse the relation to predictive fit. RJN characterizes this relative mapping variation through its amount, form and distribution over images.

1 INTRODUCTION

Two images can produce similar neural responses while engaging different combinations of visual features. An encoding model captures one version of this possibility: equal predictions at two inputs need not have equal derivatives. A feature change may strongly affect one prediction and leave the other almost unchanged. Encoding models provide a way to study how visual representations relate to brain responses [5, 20, 27]. Their fitted transformations raise a further question: how does local dependence on those features vary across images [17]?

The scale of this variation matters. A large change in a strongly sensitive mapping and a small change in a weakly sensitive mapping can represent the same proportional disagreement. Pooling their squared differences emphasizes the former; comparing each difference with the sensitivities involved gives the latter equal scope to contribute. The form of the change matters as well. Sensitivity may grow or shrink along a common direction, or the combination of features affecting the prediction may change. An overall measure becomes interpretable when these forms of variation can be distinguished and related to individual stimuli.

Existing Jacobian summaries approach this problem through different aggregations. Jacobian-based Nonlinearity Evaluation (JNE) measures dispersion of distances from the mean Jacobian, with JNE-SS adding stimulus resolution [28]. Gradient second moments describe sensitivity directions [4], and nonlinearity coefficients relate derivative energy to output variation [21, 22]. Normalized gradient-energy variation also has magnitude and directional components and a stimulus allocation. These constructions motivate a precise choice of comparison scale: should a discrepancy be normalized after averaging across the field, or relative to the two mappings being compared? This question concerns variation of the fitted transformation across inputs. Derivative-based feature attribution instead assigns an output or output difference to features [24, 25].

We introduce *relative Jacobian nonlinearity* (RJN) to make the pair-relative comparison explicit. Consider how the same small feature perturbation changes the prediction around two images. RJN expresses the disagreement between these linearized responses relative to their combined root-mean-square sensitivity, then averages over image pairs. Its kernel is the squared triangular-ratio distance on nonzero vectors [13]. The construction connects three levels of description: how much local

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.

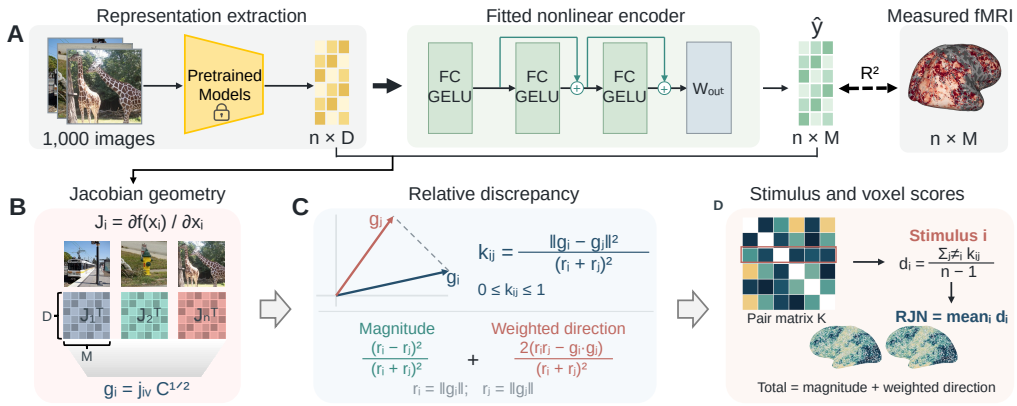

Figure 1: **From fitted encoders to relative Jacobian nonlinearity. A**, reference images yield visual features and predicted voxel responses. **B**, the full Jacobian $J_i \in \mathbb{R}^{M \times D}$ is displayed transposed; its voxel-v row j_{iv} gives $g_i = j_{iv}C^{1/2}$. **C**, the relative pair discrepancy separates into magnitude and weighted direction. **D**, voxelwise pair matrices and their stimulus and cortical summaries. Matrices and the fMRI rendering are schematic; the RJN maps show S5/CLIP/layer 10. The input geometry C and zero-gradient cases are specified in Section 2.

mappings vary, whether their disagreement concerns magnitude or weighted direction, and how that variation is distributed over images. We characterize its scaling and coordinate properties and derive its exact relation to energy weighting.

We use this description to examine visual encoders fitted to the Natural Scenes Dataset. Regional analysis asks what form a mapping contrast takes; stimulus analysis asks how mapping profiles relate to response levels; layer analysis asks how their relation to predictive fit depends on feature perturbations. Comparisons with energy variation reveal where the weighting changes the description, and comparisons across participants and fitted heads assess which image orderings recur. Together, the metric and these analyses provide a way to interpret the changing local use of a visual representation by a neural encoding model.

2 Relative Jacobian Nonlinearity

2.1 Comparing responses to common feature changes

Let $f_v(x)$ be a fitted encoding output for voxel v, with feature vector x and row Jacobian $j_v(x) = \partial f_v(x)/\partial x$. We evaluate the derivative at $n \geq 2$ reference stimuli, with the representation and fitted parameters fixed (Figure 1). The Jacobian describes the local linear response to a feature change. To compare two such mappings, we first specify the changes of interest through a positive-semidefinite geometry C, shared by all reference stimuli. The weighted Jacobian is $g_v(x) = j_v(x)C^{1/2}$, using the symmetric positive-semidefinite square root.

For a common zero-mean perturbation δ with covariance C, $\|g_v(x)\|^2$ is the expected squared linearized response. Likewise, $\|g_v(x_i) - g_v(x_j)\|^2$ is the expected squared disagreement between the two responses to that same perturbation. Identity geometry gives equal weight to feature coordinates; covariance geometry weights their variation in a stated stimulus distribution. With nonsingular covariance, g is the derivative in whitened coordinates [15]; a singular covariance restricts the comparison to its supported subspace. An invertible affine feature change preserves the score when the geometry is transformed accordingly (Appendices B.3 and B.5). Geometry is therefore chosen from the intended perturbation question. A separate feature sample can estimate covariance, and the same geometry rule is retained across compared models. The comparison also needs a response scale. Each mapping's norm is its root-mean-square sensitivity to the chosen perturbations; the sum of these sensitivities bounds the root-mean-square disagreement between the mappings. We use this bound to express disagreement relative to the sensitivities involved.

2.2 PAIR-RELATIVE DISAGREEMENT AND ITS COMPONENTS

For weighted Jacobians g, h with norms r, s, define

$$k(g,h) = \frac{\|g-h\|^2}{(r+s)^2}, \qquad k_{\mathrm{mag}}(g,h) = \frac{(r-s)^2}{(r+s)^2}, \qquad k_{\mathrm{dir}}(g,h) = \frac{2(rs - g\cdot h)}{(r+s)^2}. \tag{1}$$

Expansion gives $k = k_{\mathrm{mag}} + k_{\mathrm{dir}}$; the triangle and Cauchy–Schwarz inequalities put all three quantities in $[0,1]$. For nonzero vectors, k is the squared triangular-ratio distance. This square is a bounded dissimilarity and does not satisfy the metric triangle inequality (Appendix B.4).

The magnitude term measures unequal sensitivity norms. For positive norms, direction contributes $2rs(1-\cos\theta)/(r+s)^2$, where θ is the angle between the mappings. Its weight is largest at equal norms and decreases as they become unequal. Positively aligned mappings consequently disagree only in magnitude; equal-norm mappings disagree only in direction. The components describe two forms of change within the same comparison, so their contributions can be added across pairs and stimuli. Direction here concerns geometry-weighted feature sensitivity.

When exactly one vector is zero, the total, magnitude and directional values are $(1,1,0)$; when both are zero, all three are assigned zero. A common nonzero output gain leaves the comparison unchanged, and an intercept leaves the derivatives unchanged. An affine head has zero discrepancy. On a finite reference set, zero RJN means agreement of the weighted mappings at those inputs. Appendix B.1 gives the support and regularity conditions for extending this statement to a domain. The same relative scaling makes errors consequential when they are large compared with $r+s$; Appendix B.2 derives the perturbation bound, including the double-zero boundary.

2.3 FROM MAPPING PAIRS TO IMAGE SCORES

Writing $g_{iv} = g_v(x_i)$, define the global and stimulus scores as

$$\widehat{\mathcal{N}}_J(v) = \frac{1}{n(n-1)} \sum_{i \neq j} k(g_{iv}, g_{jv}), \qquad d_{iv} = \frac{1}{n-1} \sum_{j \neq i} k(g_{iv}, g_{jv}). \tag{2}$$

The same averages applied to each component give

$$\widehat{\mathcal{N}}_J = \widehat{\mathcal{N}}_{\mathrm{mag}} + \widehat{\mathcal{N}}_{\mathrm{dir}} = \frac{1}{n} \sum_i d_i. \tag{3}$$

Each image score describes how its local mapping differs from the reference mappings. Its mean recovers the global score, and the identity also holds componentwise and after averaging a common voxel population. Thus a regional summary and its stimulus profile describe the same pair comparisons at different resolutions. Appendix B.5 derives scoring against a separate reference; Appendices B.6 and B.7 give sampling and replacement properties for fixed geometry.

The weighting can be compared directly with normalized gradient-energy variation, $V_C = 1 - \|\mathbb{E}_n g\|^2 / \mathbb{E}_n \|g\|^2$, where $\mathbb{E}_n$ averages reference inputs and total energy is positive. Both measures support component decompositions and stimulus allocations. V_C normalizes the average squared discrepancy by total field energy; RJN normalizes each discrepancy by its pair's combined sensitivity before averaging. They are proportional for equal positive Jacobian norms. For unequal norms, their exact relation includes a norm-dependent factor and a correction for covariance between pair scale and relative disagreement (Appendix B.9). Analytic fields with matched predictions and specified scalar sensitivity summaries illustrate this distinction in Appendix B.8.

3 EVALUATION IN VISUAL ENCODING MODELS

We study four NSD participants, S1, S2, S5 and S7, using CLIP and supervised ViT-B/16 features from all twelve transformer layers [2, 8, 23]. Each layer supplies a nonlinear encoding head with a linear voxel readout, fitted and selected on separate training and validation images. Shared held-out images provide the reference mappings and measured responses. Appendix C.1 specifies the architecture, preprocessing and anatomical alignment.

The primary geometry is centered reference-feature covariance within each participant–representation–layer condition. Identity comparisons hold functions, images and outputs fixed. Sensitivity strength is $T_C(v) = n^{-1}\sum_i \|g_{iv}\|^2$, and predictive fit is held-out R^2. Analyses cover V1–V4, EBA, PPA, RSC, OPA, FFA-1 and FFA-2, averaging valid outputs within each ROI and model conditions equally. Pearson correlations compare outputs within a condition and ROI, or layers within a participant, representation and ROI. Stimulus profiles average aligned output scores and responses within each ROI and use Spearman correlations across images. All distinct reference pairs enter RJN. Appendices C.2, C.3 and C.4 give the calculation and estimation checks. The conditions are repeated model comparisons for four participants.

4 Results

4.1 What changes across cortical regions?

A regional difference in RJN can arise because sensitivities change in strength, because their weighted directions differ, or through both. The decomposition distinguishes these possibilities while retaining the pair comparisons underlying the total score.

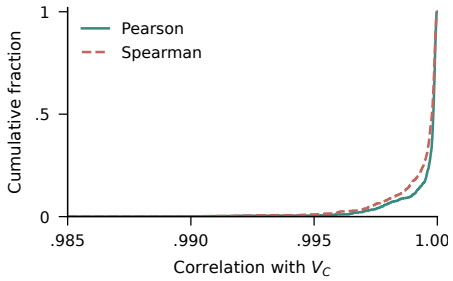

Figure 2: **RJN and gradient variation.** Cumulative distributions of within-ROI voxelwise correlations across model conditions.

RJN closely follows normalized gradient-energy variation across the evaluated models (Figure 2). Mean within-cell Pearson and Spearman correlations both exceed 0.999, with minima of 0.990 and 0.987. Their magnitude and directional components also have near-identical voxel orderings (Appendix F). Equation 9 gives the exact relation between their weightings.

Figure 3 shows the cortical and regional scores. The ten regional means average to 0.057, with direction accounting for 89.8% of the aggregate mean. Regional directional fractions range from 86.3% in OPA to 93.4% in V1. Magnitude-weighted directional disagreement therefore contributes most of the measured variation across these regions.

Table 1: Regional means over 96 conditions. Dir.: $100\,\widehat{\mathcal{N}}_{\mathrm{dir}}/\widehat{\mathcal{N}}_J$; T_C: squared-response units.

ROI	RJN	Dir. (%)	T_C	R^2
V1	0.079	93.4	0.037	0.047
V2	0.059	91.3	0.053	0.080
V3	0.061	91.2	0.049	0.075
V4	0.046	88.9	0.070	0.113
EBA	0.052	88.0	0.089	0.135
PPA	0.051	87.1	0.092	0.158
RSC	0.058	89.2	0.069	0.106
OPA	0.045	86.3	0.083	0.144
FFA-1	0.050	87.5	0.075	0.124
FFA-2	0.069	91.2	0.064	0.095

Relative variation gives a different regional ordering from sensitivity energy (Table 1). V1 has higher RJN than EBA (0.079 versus 0.052) and lower energy (0.037 versus 0.089), a contrast present in all four participant averages. Across the ten regional means, energy and RJN have Spearman correlation -0.77, with 36 of 45 ROI pairs ordered oppositely. Omitting one participant at a time leaves 30–36 discordant pairs and correlations from -0.79 to -0.58. Modest absolute sensitivity can thus accompany substantial relative mapping variation.

The contrast between the two fusiform regions illustrates the directional case. FFA-2 has higher mean RJN than FFA-1 in every participant, with differences of 0.012–0.037 and an overall increase of 39.8%. Total and directional differences are positive in all 96 conditions: direction carries the mean increase of 0.020, while the magnitude difference is -6.0×10^{-5}. The energy baseline recovers the same FFA ordering throughout the balanced comparison subset (Appendix F). The decomposition identifies the form of this shared regional pattern. Directional disagreement mostly involves shifts among mappings with nonnegative inner products: negative-inner-product pairs average 0.021% of pairs and account for at most 0.41% of aggregate directional RJN (Appendix D.1). Most directional variation therefore arises among nonzero mappings separated by at most a right angle.

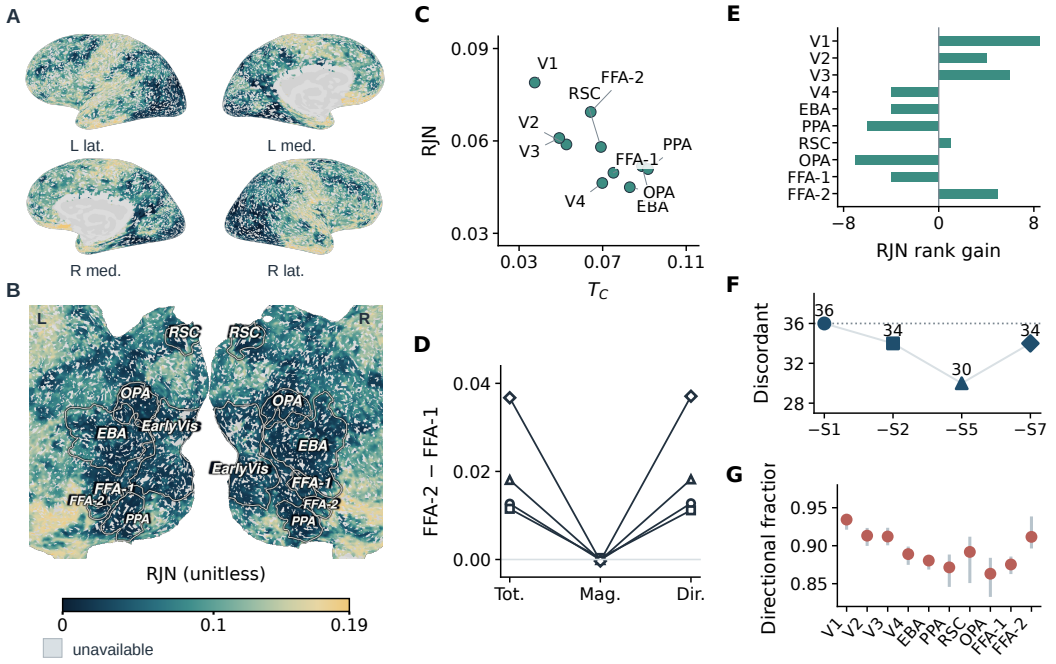

Figure 3: **Regional organization of relative Jacobian nonlinearity. A**, S5/CLIP/layer-10 cortical surfaces: left lateral and left medial above right medial and right lateral. **B**, the corresponding flatmap, with a common 0–0.19 RJN scale. Native ROI contours identify the regions; EarlyVis groups V1–V4 in cortical displays. **C**, mean sensitivity energy versus RJN for ten ROIs. **D**, participant-average FFA-2 minus FFA-1 component differences. **E**, regional rank changes from energy to RJN. **F**, rank discordances after omitting one participant. **G**, directional fractions, with segments spanning participant means. Circle, square, triangle and diamond identify S1, S2, S5 and S7. Cortical smoothing and missing support are described in Appendix J.4.

How much of the FFA contrast remains when the compared outputs have similar reliability, predictive fit and sensitivity strength? Within shared strata of these quantities, the equal-participant mean difference is 0.0048, predominantly directional. Shared support is smaller, particularly for S7; reliability-only adjustment gives a near-zero S2 contrast. Removing the bottom energy quintile of each condition's ROI union gives a mean difference of 0.0060, positive in 92 of 96 conditions, with the same sign count for energy variation. The weakest-sensitivity quintile also has greater relative gradient drift and less consistent score ranks across refits than the strongest. Changing the treatment of weak pairs has a larger effect: flooring combined sensitivity at twice the ROI union's median RMS sensitivity reverses the contrast in six of eight layer-10 comparisons on fixed output subsets. The regional contrast therefore combines a directional pattern with dependence on the output population and the sensitivity scale used in comparison. Appendices D.2 and I give the stratification, refit and modified-kernel analyses.

4.2 Which images carry the mapping variation?

A regional average leaves open how variation is distributed over stimuli. Image scores retain this information and separate changes in sensitivity strength from changes in weighted direction. These components can relate differently to recorded responses (Figure 4; Table 2). In OPA, mean response correlations are $+0.048$ for magnitude, -0.085 for direction and -0.046 for total RJN. Component signs oppose in all four participant averages and 66 of 96 conditions; PPA shows the pattern in three participants and 67 conditions.

Figure 5 makes the distinction between response level and mapping profile concrete. The six image pairs have target-ROI recorded responses within 0.0055 standardized units, yet directional scores differ by 0.0067–0.0164. Their predicted responses differ by 0.150–0.321. The low- and high-score images also occupy different positions across regions: mean directional percentiles span 0.7–25.8 and

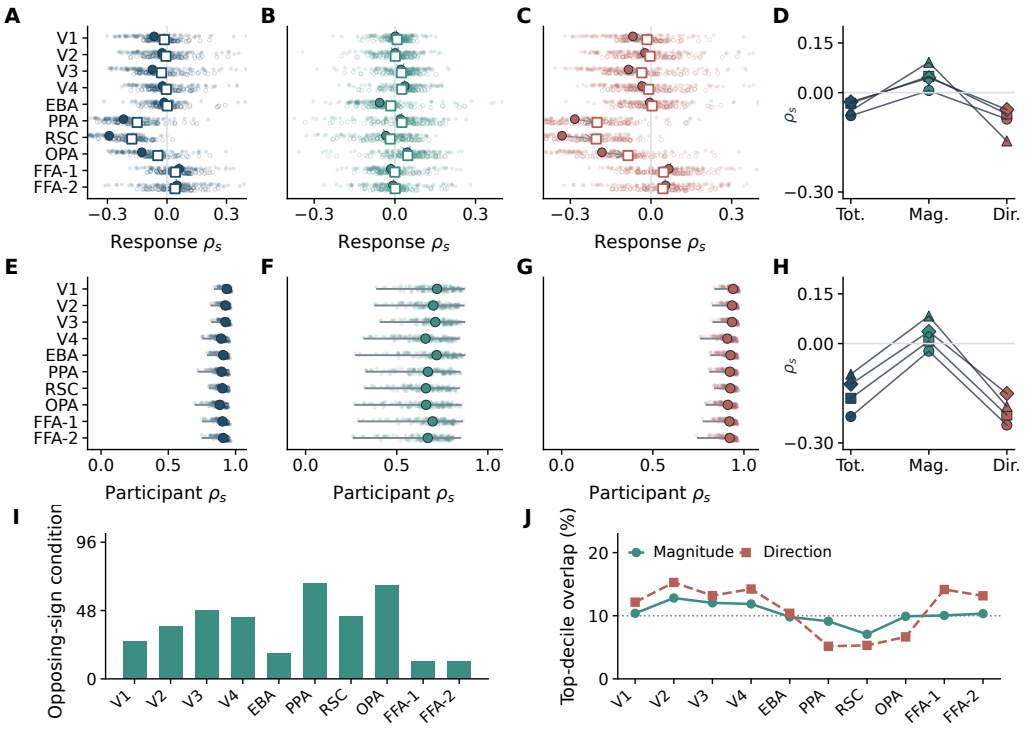

Figure 4: **Response associations and participant consistency of stimulus scores. A–C**, total, magnitude and directional scores: condition-level Spearman correlations with predicted responses (small filled points) and recorded responses (small open points). Large filled circles and open squares mark their respective means; arrows run from predicted to recorded means. **D/H**, OPA/PPA participant means of recorded-response correlations across representations and layers. **E–G**, the corresponding score-profile correlations across participants, using aligned images. Points show matched participant comparisons within each ROI; overlaid marks give their mean and range. **I**, condition counts with positive magnitude and negative directional recorded-response correlations. **J**, overlap between the top 100 score-ranked images and top 100 recorded responses, averaged over conditions. The dotted line marks the 10% expectation for independent selections. Circle, square, triangle and diamond in D/H identify S1, S2, S5 and S7. All panels use the 1,000 shared reference images.

Table 2: Mean recorded-response correlations. Opp.: conditions with positive magnitude and negative direction. W_{dir}: within-bin/global IQR ratio.

ROI	Total	Mag.	Dir.	Opp.	W_{dir}
OPA	-0.046	+0.048	-0.085	66/96	0.96
PPA	-0.150	+0.026	-0.201	67/96	0.94
RSC	-0.177	-0.017	-0.204	44/96	0.92
FFA-1	+0.043	0.000	+0.047	12/96	0.95
FFA-2	+0.041	+0.001	+0.044	12/96	0.95

Directional scores also retain substantial spread within recorded-response strata. Dividing the images into twenty equal-count response-rank bins, we compare the median within-bin directional-score IQR with the global IQR. This ratio averages 0.96 in OPA and 0.94 in PPA. Typical within-bin spread is consequently close to overall spread (Appendix D.3; Figure S4). The measure characterizes dispersion within response ranks, whose absolute response ranges can vary.

73.3–96.5, respectively. Matching the target-ROI average leaves differences elsewhere in the response pattern, with cross-ROI mean absolute recorded-response differences of 0.114–0.335 (Appendix J.1).

Similarity can also be required of both recorded and predicted responses. A tolerance of 0.01 for each yields five OPA and seven PPA pairs within the original candidate bands, with directional differences of 0.0070–0.0181. Across participants, representations and layers, joint matches occur in 187 of 192 ROI–condition searches: 908 of 84,816 candidate pairs, with a median directional difference of 0.0161 (interquartile range 0.0119–0.0238). Selecting low and high score quartiles supplies the directional separation; finding pairs within both response tolerances shows that this separation can coexist with similar measured and predicted ROI averages. Appendices J.2 and J.3 detail the selections.

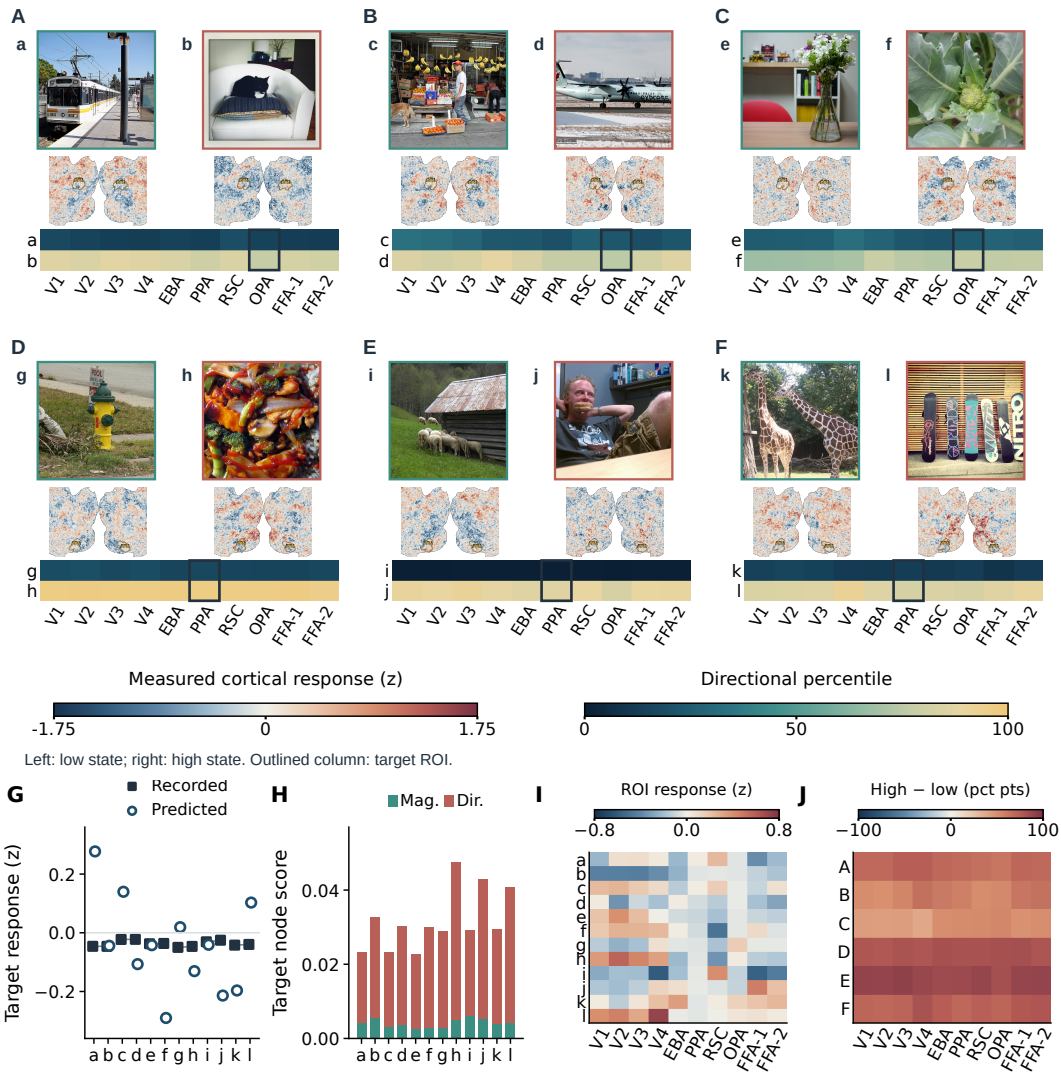

Figure 5: **Similar recorded responses with different local mappings. A–F**, six low/high directional-score pairs from S5/CLIP/layer 10: three selected in OPA and three in PPA. Each panel contains the NSD images, measured cortical responses on a shared $[-1.75, 1.75]$ scale, and a two-image-by-ten-ROI directional-percentile heatmap. Black outlines mark the selection ROI; teal and coral identify low and high states. Lowercase a–l follow the image order. **G**, recorded and predicted target-ROI responses. **H**, magnitude and directional node scores, stacked to give total scores. **I**, recorded ten-ROI response profiles. **J**, high-minus-low directional percentiles in percentage points, with rows A–F matching the pair panels. Stimuli follow the NSD presentations of MS COCO images [2, 18].

The image ordering recurs across participant models with shared images and representations. Aligned profiles have mean Spearman correlations of 0.91 for total RJN, 0.92 for direction and 0.69 for magnitude. They select different images from response rankings: top-100 directional-score and recorded-response sets overlap by 6.6% in OPA and 5.2% in PPA, compared with magnitude-score overlaps of 9.9% and 9.1% (Appendix D.3).

Image profiles also reveal the effect of the comparison scale. Although RJN and normalized energy variation nearly agree on voxel ordering, their stimulus allocations differ. Across 240 ROI–condition profiles spanning early, middle and final layers, total RJN and energy-node profiles have median Spearman correlation 0.921 and top-100 overlap 0.84; directional values are 0.924 and 0.81. Original JNE-SS has near-zero median rank correlation with total RJN despite a median top-100 overlap

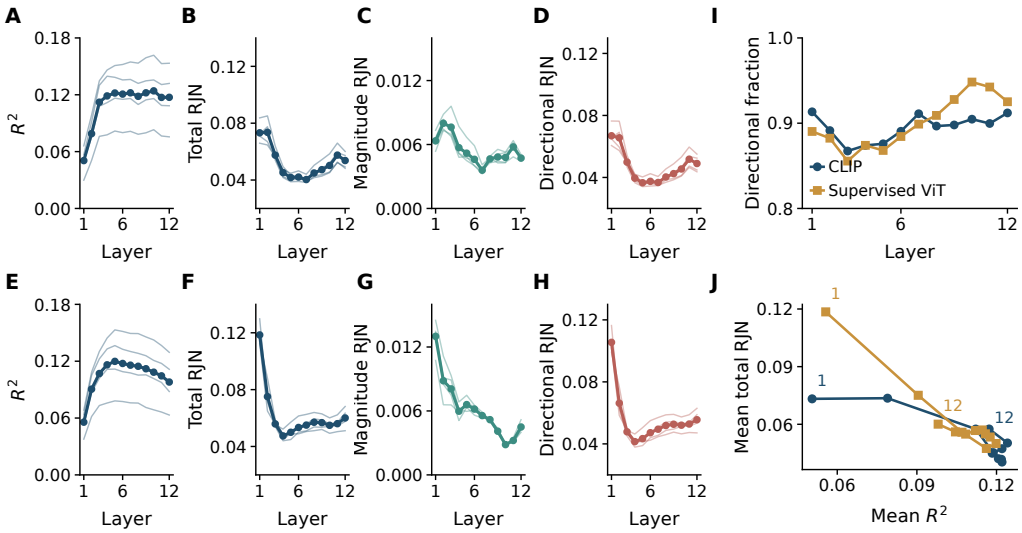

Figure 6: **Prediction and mapping variation across feature depth. A–D**, CLIP; **E–H**, supervised ViT. Columns show predictive R^2, total RJN, magnitude RJN and directional RJN, with matching vertical ranges across representations. Thin lines show participant means over ten ROIs; thick lines average participants. **I**, mean directional RJN divided by mean total RJN. **J**, mean R^2 and total RJN connected in layer order, with endpoint layers labeled. Blue circles denote CLIP and gold squares supervised ViT.

of 70%: its squared centered-distance score assigns high values to distances at either extreme (Appendix F).

Do these image orderings persist when the encoding head is fitted again? Across alternative layer-10 fits, median profile correlations are 0.931 for total RJN, 0.944 for direction and 0.664 for magnitude; magnitude voxel ranks are less consistent (median 0.355). Total energy-node profiles have median correlation 0.898 on identical output subsets, while both total measures have median top-100 overlap 0.76. Total RJN has higher profile correlation in every paired comparison, with overlap differences of both signs. The total pair-relative profiles recur more consistently by rank correlation across these fits, while their most highly ranked image sets have similar overlap (Appendices G and I).

4.3 Which feature changes define the relation to fit?

Do the layers that predict responses best also have the least variation in their local mappings? Under covariance geometry, fit and RJN have negative correlations across voxels in all 960 cells (median -0.72) and across layers in all 80 strata (median -0.87). Yet their extrema differ: fit peaks at layers 10 and 5 for CLIP and supervised ViT, whereas RJN is lowest at layers 7 and 4. Direction remains the larger component, with fractions of 0.87–0.91 across CLIP layers and 0.86–0.95 across supervised ViT layers (Figure 6; Appendix D.4).

This relationship changes when the same functions, images and outputs are compared under identity geometry. Layerwise correlations now include 44 positive and 36 negative values (median 0.08). All eight strata reverse sign in EBA, PPA, RSC and FFA-1; six reverse in OPA and FFA-2; none reverse in V1–V4 (Figure 7). Covariance geometry asks about perturbations shaped by feature variation; identity geometry gives equal weight to coordinate directions. A model's predictive fit is unchanged, but its mapping variation is measured against a different set of feature changes.

Within covariance geometry, predictive-fit accounting locates the negative association chiefly in response correlation and directional variation. The mean additive contributions from squared response correlation, gain-consistency penalty and mean-bias penalty are -0.679, -0.029 and $+0.003$; crossing this accounting with the RJN decomposition gives -0.002 for magnitude and -0.703 for direction. All terms share a correlation denominator (Appendix E; Figure S5). The association also depends on output reliability and sensitivity strength. Across the same 960 cells, the mean raw correlation is -0.705. Correlating residuals after regression on noise ceiling gives -0.055; adding log sensitivity energy gives $+0.465$, positive in 943 of 960 cells. Reliability-only coefficients include

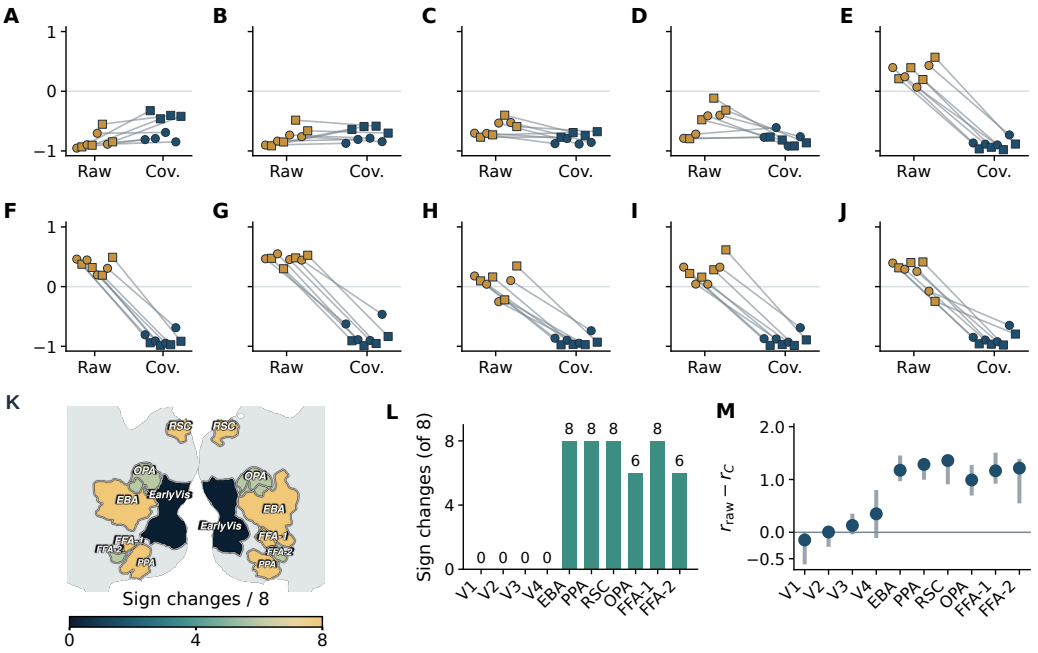

Figure 7: **Geometry changes the layerwise fit–RJN relationship. A–J**, Pearson correlations across twelve layers for V1, V2, V3, V4, EBA, PPA, RSC, OPA, FFA-1 and FFA-2. Lines connect raw-coordinate (gold) and covariance (blue) results within a participant–representation stratum; circles denote CLIP and squares supervised ViT. **K**, cortical distribution of sign-change counts. **L**, counts among the eight strata for each ROI. EarlyVis combines V1–V4. **M**, median paired correlation differences $r_{\text{raw}} - r_C$, with segments spanning the eight observed values.

both signs. Energy summarizes the same Jacobian field, so this conditioning describes the association remaining after linear adjustment for reliability and log sensitivity energy.

Geometry estimation and output selection affect different parts of this comparison. Covariance reestimation on fixed anatomical subsets changes total regional RJN by a median absolute relative change of 1.58% (Appendix G). Estimation from disjoint training images on these three-output ROI subsets preserves 79 of 80 layerwise Pearson signs, with median absolute change 0.024. The subset reference-covariance baseline has 75 negative associations, while identity gives 33 positive ones. These subsets also change the FFA comparison: the reference-covariance contrast is positive in 49 of 96 conditions. Appendix H gives both correlation coefficients and the interpolation toward isotropic geometry. Keeping the output set fixed allows the effect of the perturbation geometry to be read separately from changes in regional averaging.

5 Discussion and Conclusion

Neural encoding models are usually compared through the responses they predict. RJN complements this view with pair-relative comparisons of local feature dependence [16, 17]. Its components and image scores describe the form and stimulus distribution of variation.

Direction accounts for the fusiform contrast; images with similar ROI responses can have different mapping scores; and the relation to fit changes with the feature perturbations considered. Energy variation recovers the principal regional patterns, while pair-relative weighting changes stimulus rankings and their repeatability across the evaluated fits. Output reliability and weak sensitivity shape these descriptions, linking their interpretation to gradient estimation and the compared population.

The evidence covers two representations and one encoding-head family in NSD. Relating stimulus selections to independently measured visual properties, or composing the head with its representation to examine image-space changes, would connect these profiles to more specific visual questions.

AI USE STATEMENT

LLMs were used only to improve the writing quality of this manuscript, including grammar correction and language polishing. They were not used for generating research ideas, designing methods, conducting experiments, or analyzing results. The authors reviewed and take full responsibility for all content in the paper.

REFERENCES

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, pp. 9505–9515, 2018. URL https://papers.nips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

[2] 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*, 25:116–126, 2022. doi: 10.1038/s41593-021-00962-x. URL https://www.nature.com/articles/s41593-021-00962-x.

[3] Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153):1–43, 2018. URL https://jmlr.org/papers/v18/17-468.html.

[4] Paul G. Constantine and David F. Gleich. Computing active subspaces with Monte Carlo, 2015. URL https://arxiv.org/abs/1408.0545.

[5] 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. URL https://www.nature.com/articles/s41467-024-53147-y.

[6] Jamie D. Costabile, Kaarthik A. Balakrishnan, Sina Schwinn, and Martin Haesemeyer. Model discovery to link neural activity to behavioral tasks. *eLife*, 12:e83289, 2023. doi: 10.7554/eLife.83289. URL https://elifesciences.org/articles/83289.

[7] Jörn Diedrichsen and Nikolaus Kriegeskorte. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, 13(4):e1005508, 2017. doi: 10.1371/journal.pcbi.1005508. URL https://doi.org/10.1371/journal.pcbi.1005508.

[8] 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. URL https://openreview.net/forum?id=YicbFdNTTy.

[9] James S. Gao, Alexander G. Huth, Mark D. Lescroart, and Jack L. Gallant. Pycortex: An interactive surface visualizer for fMRI. *Frontiers in Neuroinformatics*, 9:23, 2015. doi: 10.3389/fninf.2015.00023. URL https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2015.00023/full.

[10] Xiaohui Gao, Yue Cheng, Peiyang Li, Yijie Niu, Yifan Ren, Yiheng Liu, Haiyang Sun, Zhuoyi Li, Weiwei Xing, and Xintao Hu. LinBridge: A learnable framework for interpreting nonlinear neural encoding models, 2024. URL https://arxiv.org/abs/2410.20053.

[11] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs), 2016. URL https://arxiv.org/abs/1606.08415.

[12] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196. URL https://doi.org/10.1214/aoms/1177730196.

[13] Sami Hokuni, Riku Klén, Yaxiang Li, and Matti Vuorinen. Balls in the triangular ratio metric. In *Complex Analysis and Dynamical Systems VI, Part 2*, volume 667 of *Contemporary Mathematics*, pp. 105–123. American Mathematical Society, 2016. URL https://arxiv.org/abs/1212.2331.

[14] Kendrick N. Kay, Jonathan Winawer, Aviv Mezer, and Brian A. Wandell. Compressive spatial summation in human visual cortex. *Journal of Neurophysiology*, 110(2):481–494, 2013. doi: 10.1152/jn.00105.2013. URL https://journals.physiology.org/doi/full/10.1152/jn.00105.2013.

[15] Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, 2018. doi: 10.1080/00031305.2016.1277159. URL https://doi.org/10.1080/00031305.2016.1277159.

[16] Lada Kohoutová, Juyeon Heo, Sungmin Cha, Sungwoo Lee, Taesup Moon, Tor D. Wager, and Choong-Wan Woo. Toward a unified framework for interpreting machine-learning models in neuroimaging. *Nature Protocols*, 15:1399–1435, 2020. doi: 10.1038/s41596-019-0289-5.

[17] Nikolaus Kriegeskorte and Pamela K. Douglas. Interpreting encoding and decoding models. *Current Opinion in Neurobiology*, 55:167–179, 2019. doi: 10.1016/j.conb.2019.04.002. URL https://arxiv.org/abs/1812.00278.

[18] 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. URL https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48.

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[20] 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. URL https://doi.org/10.1016/j.neuroimage.2010.07.073.

[21] George Philipp. The nonlinearity coefficient—a practical guide to neural architecture design, 2021. URL https://arxiv.org/abs/2105.12210.

[22] George Philipp and Jaime G. Carbonell. The nonlinearity coefficient—predicting generalization in deep neural networks, 2019. URL https://arxiv.org/abs/1806.00179.

[23] 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, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[24] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2013. URL https://arxiv.org/abs/1312.6034.

[25] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

[26] Mariya Toneva, Jennifer Williams, Anand Bollu, Christoph Dann, and Leila Wehbe. Same cause; different effects in the brain. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 787–825. PMLR, 2022. URL https://proceedings.mlr.press/v177/toneva22a.html.

[27] 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. URL https://doi.org/10.1073/pnas.1403112111.

[28] Haoran Yang, Cheng Yue, Mengfei Zuo, Yiheng Liu, Peiyang Li, and Xiaohui Gao. Jacobian-based interpretation of nonlinear neural encoding model. In *Advances in Neural Information Processing Systems*, volume 38, pp. 4176–4211, 2025. doi: 10.52202/085713-0148. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/065e259a1d2d955e63b99aac6a3a3081-Abstract-Conference.html.

[29] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I. Inouye, and Pradeep K. Ravikumar. On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems*, volume 32, pp. 10967–10978, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/a7471fdc77b3435276507cc8f2dc2569-Abstract.html.

[30] Alicia Zeng and Jack Gallant. Disentangling superpositions: Interpretable brain encoding model with sparse concept atoms. In *Advances in Neural Information Processing Systems*, volume 38, pp. 95964–95994, 2025. doi: 10.52202/085713-3208. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8a1bc250c372e0310d3d50892002b69b-Abstract-Conference.html.

A Related research

A.1 Which part of an encoding model is being described?

The distinction between a feature representation and its fitted response mapping motivates the present question. Visual encoding models associate image features with measured neural responses [20]. Choosing the representation defines the inputs to that mapping: task-optimized visual models provide candidate feature spaces for the comparison [27]. Comparisons across pretrained representations ask which features align with brain activity [5], while concept-based encoding models expose interpretable feature structure [30]. Regional analyses examine how representational relationships vary across cortex [26]. Nonlinear spatial summation provides an example of an explicit response transformation within an encoding model [14]. Encoding, pattern-component, and representational-similarity analyses provide related ways to compare candidate representational models [7]. Together, these approaches motivate separate questions about the supplied representation and the fitted mapping from that representation to responses [17]. RJN concerns the latter: it measures variation in the local derivatives of a fixed feature-to-response function across a specified reference distribution.

A.2 How should a field of local sensitivities be summarized?

A Jacobian field contains both sensitivity strength and changes in the local mapping. Jacobian-based Nonlinearity Evaluation (JNE) measures across-stimulus dispersion of distances from the mean Jacobian; JNE-SS adds stimulus-specific resolution [28]. Taylor-based model identification resolves receptive-field, nonlinear, and interaction terms [6]. LinBridge separates shared mapping structure from sample-selective deviations [10]. RJN compares pairs of local Jacobians directly. Its normalization makes the discrepancy relative to the two sensitivity magnitudes, and its averaging rule links the overall score to magnitude, directional, and stimulus summaries. Appendix B.8 provides analytic comparisons with JNE in which the relevant function values and scalar sensitivity summaries are explicitly matched.

The normalization determines which changes dominate a summary. Gradient second moments and active-subspace methods describe directions along which a function varies [4]. Covariance-weighted nonlinearity coefficients relate derivative energy to output variation [21, 22]. The quantities differ in their normalization and aggregation: sensitivity energy retains absolute derivative scale; an energy ratio compares a mean field with total energy; RJN normalizes each pair before averaging. Equation 9

gives the exact connection to normalized gradient-energy variation. The analytic comparisons in Appendix B.8 match specified scalar summaries, while allowing the full gradient second-moment matrices to differ.

A.3 What does a relative mapping score explain?

RJN builds a mapping comparison from an established relative distance. For nonzero vectors, the RJN pair kernel is the square of the triangular ratio distance on punctured Euclidean space [13]. Appendix B develops its zero-vector convention, radial–angular decomposition, and consequences for a geometry-weighted Jacobian field. The established distance supplies the pair comparison; the present construction uses it to connect reference-wide variation with conserved stimulus and component scores.

Interpreting a mapping score also requires identifying the object it describes. Attribution addresses how an output or output difference is assigned to input features. Gradient saliency uses a first-order approximation at the input [24]. Integrated gradients averages derivatives along an input path [25]; sanity and fidelity tests examine whether explanations respond to the model and represent its behavior [1, 29]. RJN instead compares local mappings at different reference inputs. A stimulus score measures its discrepancy from the other mappings; it does not allocate a predicted response over image pixels. The measured cortical responses accompanying the image examples provide response context, whereas the ROI score heatmaps display model-derived quantities. This distinction follows the broader separation between predictive performance, model behavior, and biological evidence in neuroimaging interpretation [16].

B Mathematical properties and analytic comparisons

B.1 Bounds, decomposition and zero-gradient cases

Expansion of the squared difference gives

$$\|g - h\|^2 = (r - s)^2 + 2(rs - g \cdot h).$$

Cauchy–Schwarz makes both terms nonnegative. The triangle inequality gives $\|g - h\| \le r + s$, proving the unit bound whenever the denominator is positive. Averaging preserves the component decomposition and all three bounds.

The double-zero convention extends the relative comparison to a silent pair. This extension is discontinuous: for nonzero u and $t \neq 0$, $k(tu, -tu) = 1$, whereas $k(0, 0) = 0$. As one vector tends to zero while the other remains nonzero, $(k, k_{\mathrm{mag}}, k_{\mathrm{dir}})$ tends to $(1, 1, 0)$. Computation must therefore preserve the distinction between exact zero and small nonzero vectors. A common positive rescaling of the compared vectors preserves the ratio and controls numerical range. The numerical evaluation below uses one such scale per output across its reference field.

An empirical average of nonnegative pair discrepancies vanishes exactly when all weighted Jacobians agree. For independent draws from a population distribution with fixed geometry, zero expected discrepancy likewise implies a common weighted Jacobian almost surely. Extending this conclusion to an affine function on a domain requires support and regularity: a continuously differentiable function with constant gradient throughout a connected open region is affine there. Finite reference data establish agreement on their sampled points, and singular geometry restricts the measured directions.

B.2 Perturbation size relative to sensitivity

The relevant error scale for a relative discrepancy is the perturbation size divided by the pair's sensitivity. Write $g' = g + e_g$, $h' = h + e_h$, $S = \|g\| + \|h\|$ and $\epsilon = \|e_g\| + \|e_h\|$. For a fixed nonnegative floor τ, define $D_\tau = \max(S, 2\tau)$ and $k_\tau(g, h) = \|g - h\|^2 / D_\tau^2$, with the same double-zero convention. At $\tau = 0$ this is the original kernel. Whenever $D_\tau > 0$,

$$|k_\tau(g', h') - k_\tau(g, h)| \le \min\{1, 4\epsilon / D_\tau\}.$$

To see this, let $a = \|g - h\|$ and use primes for the perturbed quantities. The norm and maximum are Lipschitz, so $|a' - a| \le \epsilon$ and $|D'_\tau - D_\tau| \le \epsilon$. When $D'_\tau > 0$, both normalized distances are at most one, and

$$\left| \frac{a'}{D'_\tau} - \frac{a}{D_\tau} \right| \le \frac{|a' - a|}{D_\tau} + \frac{a'}{D'_\tau} \frac{|D_\tau - D'_\tau|}{D_\tau} \le \frac{2\epsilon}{D_\tau}.$$

Factoring the difference of squares proves the bound. If $D'_\tau = 0$, then $\tau = 0$ and $\epsilon \geq S$, so the unit bound applies. Replacing a by $|\,\|g\| - \|h\|\,|$ gives the same bound for the magnitude component; additivity gives $\min\{1, 8\epsilon/D_\tau\}$ for direction. For $\tau > 0$, the total bound is at most $\min\{1, 2\epsilon/\tau\}$, including an originally double-zero pair. A positive floor thus changes the diagnostic's treatment of weak pairs and supplies an absolute perturbation scale. The unfloored score instead requires errors small relative to S; small absolute derivative error alone does not ensure a small score error. Appendix I evaluates output filtering, floors and observed changes between fitted heads.

B.3 Scaling and coordinate changes

For an output transformation $f \mapsto cf + d$, $c \neq 0$, both the numerator and denominator scale by c^2. An additive intercept has no derivative. Reference permutation only reorders the pair terms.

For an invertible feature transformation $z = Bx+b$, define $\widetilde{f}(z) = f(B^{-1}(z-b))$ and $C_z = BC_xB^\top$. At corresponding inputs,

$$j_{\widetilde{f}}(z_i) = j_f(x_i)B^{-1}, \qquad j_{\widetilde{f}}(z_i)C_z j_{\widetilde{f}}(z_j)^\top = j_f(x_i)C_x j_f(x_j)^\top.$$

Every norm and inner product in the kernel is consequently preserved. This argument applies to positive-semidefinite geometry because it uses the quadratic forms directly. An independently added diagonal regularizer generally changes these forms across coordinate systems. Nonlinear representation changes also fall outside this affine reparameterization result.

For a singular $C_x = U\,\mathrm{diag}(\lambda_a)U^\top$, with $\lambda_a \geq 0$, the coordinates $z_a = \lambda_a^{-1/2}u_a^\top(x - \bar{x})$ are defined only where $\lambda_a > 0$. On this supported subspace, $x = \bar{x} + U_+\,\mathrm{diag}(\sqrt{\lambda_a})z$ and the derivative with respect to z is $jU_+\,\mathrm{diag}(\sqrt{\lambda_a})$. Its norm and pairwise inner products equal those of $jC_x^{1/2}$. Null-space directions have zero weight rather than an inferred whitened coordinate; hence RJN identifies variation of the mapping only along covariance-supported perturbations.

B.4 Relationship to the triangular ratio distance

For nonzero g, h in $G = \mathbb{R}^d \setminus \{0\}$, the triangular ratio distance is

$$s_G(g,h) = \frac{\|g-h\|}{\|g\| + \|h\|}$$

[13]. The RJN pair kernel is s_G^2, extended to zero vectors as specified above. Squaring does not preserve the distance triangle inequality: for positive collinear scalars,

$$k(1,3) = \frac{1}{4} > \frac{1}{9} + \frac{1}{25} = k(1,2) + k(2,3).$$

The squared kernel is thus a bounded dissimilarity without the metric triangle inequality. Its radial–angular expansion supplies the magnitude–direction split used to interpret variation in the weighted Jacobian field.

B.5 Common perturbations and stimulus conservation

With a common zero-mean perturbation δ having covariance C,

$$\mathbb{E}(j_i\delta)^2 = j_i C j_i^\top, \qquad \mathbb{E}[(j_i - j_j)\delta]^2 = (j_i - j_j)C(j_i - j_j)^\top.$$

Writing $u_i = j_i\delta$ and $u_j = j_j\delta$, substitution yields the response identity

$$k(g_i, g_j) = \frac{\mathbb{E}(u_i - u_j)^2}{\left(\sqrt{\mathbb{E}u_i^2} + \sqrt{\mathbb{E}u_j^2}\right)^2}, \tag{4}$$

whenever combined RMS sensitivity is positive. At a double-zero pair, the response ratio takes the same assigned zero value as $k(0,0)$. Only the perturbation's second moment enters the identity. Sharing the perturbation preserves the cross term between the two local responses; independent perturbations would remove it. Application to finite function changes also requires accurate local linearization.

For symmetric pair scores, the ordered-pair and unordered-pair definitions agree:

$$\widehat{\mathcal{N}}_J = \frac{1}{n(n-1)}\sum_{i \neq j} k_{ij} = \frac{2}{n(n-1)}\sum_{i<j} k_{ij}.$$

Summing the node scores $d_i = (n-1)^{-1}\sum_{j\neq i} k_{ij}$ counts every unordered pair at each of its two endpoints. Division by n therefore gives Eq. 3 for the total and both components. The pair terms share reference inputs and are not independent observations. For an evaluation set A and a separately fixed reference set B, an external-reference score and its between-set average are

$$d(x;B) = \frac{1}{|B|}\sum_{b\in B} k(g(x), g(b)), \qquad D(A,B) = \frac{1}{|A|}\sum_{x\in A} d(x;B).$$

The same definitions apply componentwise. All evaluations against B share its geometry and reference mappings; an evaluation image need not be in B. If $A = B$, these formulas include self-pairs and use denominator $|B|$, so the off-diagonal conservation identity in Eq. 3 instead uses its stated $n-1$ node denominator.

B.6 Sampling variance for a fixed encoder and geometry

For an encoder and geometry fixed before independent, identically distributed draws $X_i \sim P$, define $G_i = j(X_i)C^{1/2}$, $a(g) = \mathbb{E}k(g, G_2)$, $\theta = \mathbb{E}k(G_1, G_2)$, $K_{ij} = k(G_i, G_j)$, $\zeta_1 = \mathrm{Var}[a(G_1)]$ and $\zeta_2 = \mathrm{Var}(K_{12})$. This pair average is an order-two U-statistic [12]. Then $\mathbb{E}\widehat{\mathcal{N}}_J = \theta$ and

$$\mathrm{Var}(\widehat{\mathcal{N}}_J) = \frac{4(n-2)}{n(n-1)}\zeta_1 + \frac{2}{n(n-1)}\zeta_2. \tag{5}$$

Bounded kernels have finite variance. Conditional independence given G_1 gives $\mathrm{Cov}(K_{12}, K_{13}) = \zeta_1$; kernels on disjoint index pairs are independent. There are $\binom{n}{2}$ individual pair terms and $3\binom{n}{3}$ unordered pairs of terms sharing exactly one endpoint. Consequently,

$$\mathrm{Var}\left[\binom{n}{2}^{-1}\sum_{i<j} K_{ij}\right] = \binom{n}{2}^{-2}\left\{\binom{n}{2}\zeta_2 + 6\binom{n}{3}\zeta_1\right\},$$

which reduces to Eq. 5. Each pair has expectation θ, proving unbiasedness. This calculation applies to a geometry fixed independently of the sampled reference; re-estimating geometry from those observations changes the joint dependence of the kernels. The empirical subsampling experiment holds the full-reference covariance fixed and measures approximation to that finite-reference score.

B.7 A replacement bound under fixed geometry

Replacing m of n reference nodes can change at most $m(2n-m-1)/2$ of the $n(n-1)/2$ unordered pairs. Since each discrepancy lies in $[0, 1]$,

$$|\widehat{\mathcal{N}}_J^{\text{replaced}} - \widehat{\mathcal{N}}_J| \leq \frac{m(2n-m-1)}{n(n-1)}. \tag{6}$$

For one replacement the bound is $2/n$, irrespective of Jacobian magnitude. The same argument holds separately for the magnitude and directional components. For the corresponding population functional with fixed kernel, point-mass influence is

$$2\{\mathbb{E}[k(g,G)] - \mathbb{E}[k(G,G')]\},$$

whose absolute value is at most two. The bound assumes a fixed geometry. Re-estimating covariance after replacing inputs can change every weighted Jacobian, so robustness of the geometry estimate requires a separate analysis.

B.8 Matched predictions and scalar sensitivity summaries

On $X = \{-1, 1\}^2$ with equal weights and $C = I$, consider

$$f_A = 2x_1 + \tfrac{1}{2}x_1^2, \qquad f_B = 2x_1 + \tfrac{1}{\sqrt{3}}x_2^2 + \tfrac{1}{2} - \tfrac{1}{\sqrt{3}}, \qquad f_C = 2x_1 + \tfrac{1}{2}x_2^2.$$

All three functions equal $2x_1 + 1/2$ on the reference set. Their gradients are $(2 + x_1, 0)$, $(2, 2x_2/\sqrt{3})$ and $(2, x_2)$, respectively. For comparison with Jacobian-based Nonlinearity Evaluation (JNE) [28], let $\bar{j} = n^{-1}\sum_i j_i$, $a_i = \|j_i - \bar{j}\|_1$ and $\bar{a} = n^{-1}\sum_i a_i$, and use

$$\mathrm{JNE} = \left\{n^{-1}\sum_i (a_i - \bar{a})^2\right\}^{1/2}.$$

Table S1: Exact summaries for the analytic functions. The scalar-output nonlinearity coefficient is $\mathrm{NLC}^2 = T/\operatorname{Var}_n(f)$, with $\operatorname{Var}_n(f) = 4$.

Function	T	V_C	NLC^2	JNE	$\widehat{\mathcal{N}}_J$	$\widehat{\mathcal{N}}_{\mathrm{mag}}$	$\widehat{\mathcal{N}}_{\mathrm{dir}}$
f_A	5	1/5	5/4	0	1/6	1/6	0
f_B	16/3	1/4	4/3	0	1/6	0	1/6
f_C	5	1/5	5/4	0	2/15	0	2/15

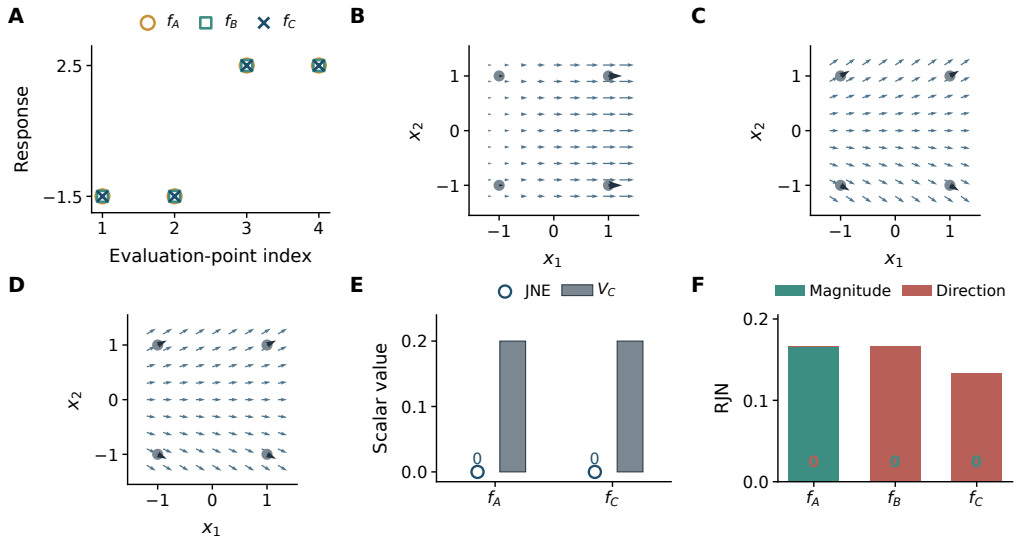

Figure S1: **Matched predictions and distinct gradient fields. A**, function values at the reference inputs $(-1,-1)$, $(-1,1)$, $(1,-1)$ and $(1,1)$, indexed 1–4. **B–D**, their analytic gradient fields on a 9×9 display grid; dark arrows identify the four reference gradients. Coordinate and arrow scales are shared. **E**, JNE and normalized gradient-energy variation V_C for f_A and f_C. **F**, RJN split into magnitude and direction. The scalar comparisons use only the four reference inputs (Table S1).

This JNE comparator is the population standard deviation, across stimuli, of L_1 distances from the mean Jacobian. For f_A and f_C, the mean is $(2, 0)$, total energy is 5, mean-field energy is 4, residual energy is 1, and $V_C = 1/5$. Constant a_i values give $\mathrm{JNE} = 0$.

For f_A, eight of the twelve ordered distinct pairs have radii 1 and 3 and a common direction. Each such pair contributes $1/4$, so $(\widehat{\mathcal{N}}_{\mathrm{mag}}, \widehat{\mathcal{N}}_{\mathrm{dir}}) = (1/6, 0)$. For f_C, all radii equal $\sqrt{5}$, and eight pairs differ by $(0, 2)$ and contribute $1/5$. Hence $(\widehat{\mathcal{N}}_{\mathrm{mag}}, \widehat{\mathcal{N}}_{\mathrm{dir}}) = (0, 2/15)$. These functions have matched prediction vectors and scalar energies but different relative variation and different component allocations. For f_B, all gradient norms are $4/\sqrt{3}$; eight pairs reverse the second component and contribute $1/4$. Thus $(\widehat{\mathcal{N}}_{\mathrm{mag}}, \widehat{\mathcal{N}}_{\mathrm{dir}}) = (0, 1/6)$. This function matches f_A in total RJN while differing in composition. It has total energy $16/3$, so this equal-total example is distinct from the matched-energy comparison between f_A and f_C.

The complete gradient second-moment matrices are $\operatorname{diag}(5, 0)$ for f_A and $\operatorname{diag}(4, 1)$ for f_C. The example therefore matches the listed scalar summaries while preserving differences in the full second moments. More generally, a balanced two-point field $g = \mu \pm a$, $a \neq 0$, has constant distance from its mean under any norm. Dispersion of those distances is zero even though direct pair discrepancy is positive.

B.9 Energy weighting identity

Let $\mathbb{E}_n$ denote an equal average over the n reference inputs. For $T = \mathbb{E}_n r_i^2 > 0$, write $m = \mathbb{E}_n r_i$ and $\mu = \mathbb{E}_n g_i$. Then $V_C = 1 - \|\mu\|^2 / T$ and $V_{\mathrm{mag}} = 1 - m^2/T$. With $q_{ij} = (r_i + r_j)^2$ and $\langle \cdot \rangle_o$

denoting the average over ordered distinct pairs, direct summation gives

$$\langle qk\rangle_o = \frac{2n}{n-1}(T - \|\mu\|^2), \tag{7}$$

$$\langle q\rangle_o = \frac{2\{(n-2)T + nm^2\}}{n-1}. \tag{8}$$

Using $\langle qk\rangle_o = \langle q\rangle_o\langle k\rangle_o + \mathrm{Cov}_o(q,k)$ and $V_{\mathrm{mag}} = 1 - m^2/T$ gives

$$\widehat{\mathcal{N}}_J = \frac{nV_C}{2(n-1) - nV_{\mathrm{mag}}} - \frac{\mathrm{Cov}_o(q,k)}{\langle q\rangle_o}. \tag{9}$$

Here $\mathrm{Cov}_o(q,k) = \langle qk\rangle_o - \langle q\rangle_o\langle k\rangle_o$ quantifies the association between pair scale and relative discrepancy. The covariance is taken over pairs with shared endpoints. If all reference radii have the same positive value, q is constant and $V_{\mathrm{mag}} = 0$, giving $\widehat{\mathcal{N}}_J = nV_C/[2(n-1)]$. Equal norms among only the nonzero nodes do not suffice when zero nodes are also present.

The energy measure has the nonnegative decomposition

$$V_C = V_{\mathrm{mag}} + V_{\mathrm{dir}}, \qquad V_{\mathrm{mag}} = \frac{T - m^2}{T}, \qquad V_{\mathrm{dir}} = \frac{m^2 - \|\mu\|^2}{T}. \tag{10}$$

Jensen's inequality gives $T \geq m^2$, and the triangle inequality gives $m \geq \|\mu\|$. Its conserved stimulus allocation is

$$d_i^V = \frac{1}{2nT}\sum_j \|g_i - g_j\|^2 = \frac{r_i^2 + T - 2g_i\cdot\mu}{2T}, \tag{11}$$

$$d_i^{V,\mathrm{mag}} = \frac{r_i^2 + T - 2r_i m}{2T}, \qquad d_i^{V,\mathrm{dir}} = \frac{r_i m - g_i\cdot\mu}{T}. \tag{12}$$

The components add at each stimulus, and their image means recover Eq. 10. These nodes average all n pairs, including the zero diagonal, whereas RJN uses $n-1$ distinct partners. Multiplying the energy nodes and their means by $n/(n-1)$ puts them on the off-diagonal convention without changing ranks. Energy nodes can exceed one even though their global means are bounded. For $T = 0$, all mappings vanish and all energy quantities are assigned zero. The distinction from RJN is the pair weighting: a common denominator $2T$ versus the pair-dependent $(r_i + r_j)^2$. Decomposition and stimulus conservation hold for both constructions.

B.10 Centered finite-difference approximation

For independent, identically distributed zero-mean probes δ_t with finite second moments and covariance C, let $z_i = (j_i\delta_t)_{t=1}^R/\sqrt{R}$. The strong law gives $z_i \cdot z_j \to j_i C j_j^\top$ simultaneously over a fixed finite collection of reference inputs. Away from double-zero pairs, continuity then gives convergence of the kernels and their averages.

For differentiable f, the centered difference

$$\frac{f(x_i + h\delta_t) - f(x_i - h\delta_t)}{2h}$$

converges to $j_i\delta_t$ as $h \to 0$ for fixed probes. Taking this limit before $R \to \infty$ establishes the sequential approximation. Finite-step error and finite-probe error can be evaluated separately. At double-zero gradients, higher-order finite-difference terms may keep a nonzero relative mismatch as their absolute magnitude vanishes, requiring separate boundary analysis. No such pair occurs among the selected outputs in the reported experiment.

C Models, data and numerical evaluation

C.1 Encoding functions and reference identities

The analysis includes NSD participants S1, S2, S5 and S7, CLIP ViT-B/16 and supervised ViT-B/16 features, and all twelve transformer layers [2, 8, 23]. These complete participant–representation–layer combinations give 96 fixed heads. The representations share architecture scale and differ in contrastive language–image versus supervised image-classification pretraining.

The model-development split used 8,000 fitting images, 1,000 validation images and 1,000 shared held-out images per participant. The present analysis evaluates the resulting fixed heads on the shared reference images. Each head maps a 768-dimensional feature vector through an affine transformation and GELU, two residual affine–GELU blocks of width 768, and a linear voxel readout:

$$h_0 = \mathrm{GELU}(W_0 x + b_0), \quad h_{\ell+1} = h_\ell + \mathrm{GELU}(W_{\ell+1} h_\ell + b_{\ell+1}),\ \ell = 0, 1, \quad \hat{y} = W_{\mathrm{out}} h_2 + b_{\mathrm{out}}.$$

The 1,000 shared images are excluded before splitting the remaining 9,000 images into 8,000 fitting and 1,000 validation images using random seed 0. The analyzed heads have width 768, biases, GELU activations and zero dropout. Each training epoch evaluates mean squared error over all 8,000 fitting images and all trained outputs, then takes one AdamW update [19] with learning rate 10^{-3}. The optimizer uses weight decay 10^{-2} and its default moment and denominator constants. Training allows up to 512 epochs and stops after 32 consecutive epochs without a lower mean validation MSE. The parameters at the minimum validation MSE are restored for evaluation. Across the 96 analyzed heads, training lasted 50–80 epochs; the selected epochs ranged from 18 to 48, counting the first epoch as one. Model development uses seed 0, separately from the anatomical sampling seed described below. The RJN analyses retain these fitted parameters and do not select a new head according to an RJN result.

Responses are standardized within acquisition runs, then averaged across repeated presentations of each image. Participant-specific rows follow sorted unique stimulus identities. Matching these identities to the official NSD design aligns feature vectors, photographs, recorded responses, and node scores. Displayed photographs are the 425-pixel-square NSD presentations with the experimental crop.

The trained output set is a fixed anatomical subset of the participant's cortical mask. Approximately half of the available outputs are sampled proportionally within each of 22 anatomical macroregions, using seed 20260508, and the union is reused across both representations and all twelve layers. The ten functional ROIs used for statistical summaries are intersections with this trained output set. Output identity is maintained through participant, masked-voxel position and native brain-mask index. The retained output counts are 50,694, 50,777, 44,690 and 42,803 for S1, S2, S5 and S7. The same identities are retained across each participant's twenty-four representation–layer conditions, giving 4,535,136 voxel–model evaluations. ROI membership may overlap. EBA denotes extrastriate body area; PPA, parahippocampal place area; RSC, retrosplenial cortex; OPA, occipital place area; and FFA-1/FFA-2, fusiform face-area subdivisions. V1–V4 remain separate statistical regions even where an EarlyVis display label groups them.

Primary summaries include ROI outputs with valid quantities. Valid voxel outputs are averaged within each participant–representation–layer–ROI cell, with the 96 conditions weighted equally in regional summaries. Pearson correlations compare outputs within 960 such cells or layers within 80 participant–representation–ROI strata. The ten ROIs are V1–V4, EBA, PPA, RSC, OPA, FFA-1 and FFA-2. Reliability-restricted summaries use noise-ceiling thresholds of 0.05 and 0.10 applied to precomputed response-derived voxelwise estimates. Missing or undefined values are omitted.

C.2 Covariance geometry and blocked Jacobian evaluation

Let $X \in \mathbb{R}^{n \times D}$ contain the reference features in the coordinates used by the fitted head. The centered empirical geometry is

$$C = \frac{(X - \mathbf{1}\bar{x}^\top)^\top (X - \mathbf{1}\bar{x}^\top)}{n}, \qquad n = 1000, \qquad C^{1/2} = U\,\mathrm{diag}(\sqrt{\max(\lambda_a, 0)})U^\top.$$

The covariance and eigendecomposition are evaluated in double precision, with denominator n and no added diagonal regularizer. An eigenvalue below $-10^{-10} \max_a |\lambda_a|$ rejects the geometry; smaller numerical negatives are set to zero. The square root is then cast to single precision for the primary Jacobian calculation. Covariance geometry is estimated separately for each participant–representation–layer condition using the full shared reference set. It weights observed feature directions; a zero eigenvalue contributes no weight. Raw-coordinate comparisons replace this operator by I while retaining the images, outputs and fitted functions.

For GELU $\phi(z) = z\Phi(z)$ [11], the derivative is $\phi'(z) = \Phi(z) + z\varphi(z)$, where Φ and φ are the standard-normal distribution and density. Writing $H_\ell = \partial h_\ell / \partial x$, the encoder derivative follows

$$H_0 = \mathrm{diag}(\phi'(W_0 x + b_0))W_0, \qquad H_{\ell+1} = H_\ell + \mathrm{diag}(\phi'(W_{\ell+1} h_\ell + b_{\ell+1}))W_{\ell+1} H_\ell.$$

Sensitivity energy is $T_C(v) = n^{-1} \sum_i \|g_{iv}\|^2$, and predictive fit is $R^2 = 1 - \sum_i (y_{iv} - \hat{y}_{iv})^2 / \sum_i (y_{iv} - \bar{y}_v)^2$. Both use the same shared reference images.

The voxel Jacobian is $j_v = W_{\mathrm{out},v} H_2$. Hidden derivatives are computed in image batches of 32 and reused across output blocks. Multiplication by $C^{1/2}$ precedes the voxel readout, which gives $g_{iv} = W_{\mathrm{out},v} H_2(x_i) C^{1/2}$ without changing the derivative. The primary evaluator uses single-precision matrix products with TensorFloat-32 disabled.

Within an output block, each reference field is divided by its largest absolute coordinate over all images and feature dimensions. An identically zero field uses scale one. The Gram matrix provides inner products, and its diagonal provides squared norms. A zero pair denominator is replaced by one only when both norms are zero; the corresponding numerators are already zero. Diagonal pair terms are set to zero. Negative directional terms from Gram subtraction are clipped to zero when their magnitude is within 2×10^{-5}; a larger negative value stops the evaluation. Magnitude and directional terms are reduced in double precision, divided by $n - 1$ for node scores and averaged over n nodes for voxel scores. Absolute sensitivity energies restore the squared output-specific scale. This calculation retains all $n(n-1)$ ordered distinct pairs rather than sampling pairs. At $n = 1000$, these are the two orientations of 499,500 unique pairs.

For a dense width-D head with a fixed number of residual blocks and M outputs, hidden derivative propagation costs $O(nD^3)$, readout projection costs $O(MnD^2)$, and the Gram comparisons cost $O(Mn^2D)$. Covariance construction and its square root cost $O(nD^2 + D^3)$. Processing B outputs at a time requires working arrays of order $O(nD^2 + BnD + Bn^2)$ beyond the model parameters and accumulated summaries. The implementation caches the $n \times D \times D$ hidden field and materializes a full $n \times n$ pair matrix only for the current output block, whose default size is 64. Thus blocking limits output-wise memory use, while exact reference-pair evaluation remains quadratic in n.

C.3 Derivative, precision and conservation checks

For the derivative check, each of the 96 model conditions uses reference positions 0, 137 and 999 and four outputs sampled evenly from the predefined prediction-significant ordering. Let J^{an} and J^{AD} collect the analytic and automatic-differentiation Jacobians for these cases [3]. We report $\max |J^{\mathrm{an}} - J^{\mathrm{AD}}|$ and the relative error $\left\|J^{\mathrm{an}} - J^{\mathrm{AD}}\right\|_2 / \max(\left\|J^{\mathrm{AD}}\right\|_2, 10^{-30})$. A separate pair check uses the first four outputs in anatomical output order and seventeen evenly spaced reference positions, evaluates $\|g_i - g_j\|^2 / (\|g_i\| + \|g_j\|)^2$ directly in double precision, and records the maximum absolute discrepancy from the matrix-based evaluation. Conservation error is the largest difference between the image-average node score and the corresponding output-level pair average. The fit-identity error is the maximum of $|R^2 - \rho^2 + (a - \rho)^2 + b^2|$ over evaluated outputs.

Across the 96 conditions, the largest analytic–automatic-differentiation error is 1.68×10^{-8} in absolute value and 4.21×10^{-7} in relative L_2 norm. The largest pair discrepancy from direct double-precision evaluation is 6.99×10^{-7}. Node conservation differs by at most 5.00×10^{-16}, and the predictive-fit identity by 1.31×10^{-15}.

A thirty-output S5/CLIP/L10 check samples three anatomically ordered outputs from each ROI, independently of RJN and R^2. Evaluating all 1,000-image pairs in double precision gives a maximum absolute difference in total RJN of 2.4×10^{-9} from the primary calculation. A separate sensitivity-energy check compares total, mean-field and residual energies for four outputs per model condition (1,152 scalar comparisons); the largest discrepancy is 7.2×10^{-9}. Stimulus scores are computed directly from the pair kernel.

C.4 Finite-reference and finite-difference estimation

Finite-reference sampling uses the same thirty ROI-balanced outputs. At each sample size 32, 64, 128, 256, and 500, fifty PCG64 draws without replacement are taken from the 1,000-image set using seed 20260919. Draws are shared across ROIs and components, with covariance held at its 1,000-image estimate. Figure S2E averages absolute errors from the 1,000-image target over ten ROIs and fifty draws. The resulting curves quantify approximation of the finite 1,000-image score with geometry held fixed. They do not include changes from re-estimating covariance or from fitting another encoder.

The finite-difference check uses one fixed S5/CLIP/L10 head, sixteen evenly spaced reference images and thirty ROI-balanced outputs. For $R \in \{64, 256, 1024\}$, three sets of shared Rademacher probes with seeds 20260919–20260921 are transformed by the symmetric $C^{1/2}$ estimated from all 1,000 reference features. Let $P(J)$ denote the total, magnitude and directional RJN values from exact projected Jacobians, $P_R(J\delta)$ the corresponding probe projection, and $P_R(\Delta_h f)$ the centered finite-difference estimate. Reconstruction error is the mean absolute difference $|P_R(\Delta_h f) - P(J)|$ over

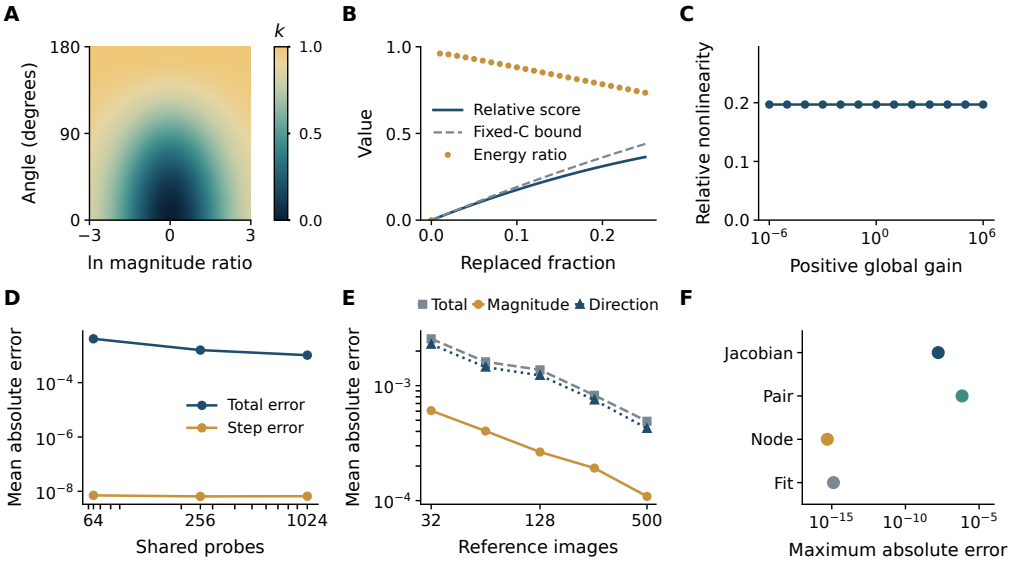

Figure S2: **Definition and estimation checks. A**, the unitless pair discrepancy k (Eq. 1) as a function of natural-log magnitude ratio and angle. **B**, replacing unit-magnitude nodes by magnitude-100 nodes: relative Jacobian nonlinearity (legend: Relative score) in blue, the fixed-geometry bound (Fixed-C bound) in dashed gray, and normalized gradient-energy variation (Energy ratio) in gold. All three are unitless. **C**, scores under positive output gains. **D**, probe-dependent centered finite-difference reconstruction error and its finite-step component at step 0.001. **E**, subsampling error relative to the 1,000-image target: total in dashed gray, magnitude in solid gold and direction in dotted blue. Panel C uses a logarithmic horizontal gain axis; D–E use logarithmic axes, and F uses a logarithmic horizontal axis. **F**, maximum errors for Jacobian, pair, node-conservation (Node) and fit-identity (Fit) checks at the evaluation scopes described in this subsection. A–C are analytic controls; D–E use thirty existing S5 outputs.

outputs and components; finite-step error replaces $P(J)$ with $P_R(J\delta)$. Figure S2D shows the seed-averaged errors at $h = 0.001$. The tested step sizes are 0.01 and 0.001. Across these step sizes, the largest finite-step component error is 3.1×10^{-6}.

D Regional, stimulus, layer and geometry results

D.1 Participant-resolved regional structure

Regional means in Figure S3 first average the 24 representation–layer conditions within each participant, then the four participants equally. A directional fraction is the ratio of mean direction to mean total. Regional ranking compares all $\binom{10}{2} = 45$ pairs of means (Table 1). Leaving out one participant recomputes each ROI mean from the remaining 72 conditions before repeating the 45 comparisons. These four counts summarize sensitivity to participant composition. FFA contrasts pair regions within each participant, representation, and layer; averaging within participant gives the four plotted summaries.

Let $p_-(v)$ be the fraction of reference pairs with $g_{iv} \cdot g_{jv} < 0$, and let $N_-(v)$ be their contribution to the directional score. Since $0 \le k_{\mathrm{dir}} \le 1$, $0 \le N_-(v) \le p_-(v)$. The bound is preserved by the output and cell averaging. Mean pair fraction is 2.1×10^{-4} and mean directional RJN is 0.051, giving $\overline{N_-}/\overline{N_{\mathrm{dir}}} < 0.0041$ from the unrounded quantities. Thus negative-inner-product pairs contribute at most 0.41% of aggregate directional RJN. The bound covers all obtuse pairwise angles among nonzero gradients, including exact reversals, while the remaining directional term still includes magnitude weighting.

D.2 FFA contrasts on shared covariate support

The original FFA comparison takes FFA-2 minus FFA-1 means within each participant–representation–layer cell, then averages the 24 cells within each participant and the four participants equally. Recomputed from the aligned output fields, the mean differences in total, magnitude and direction

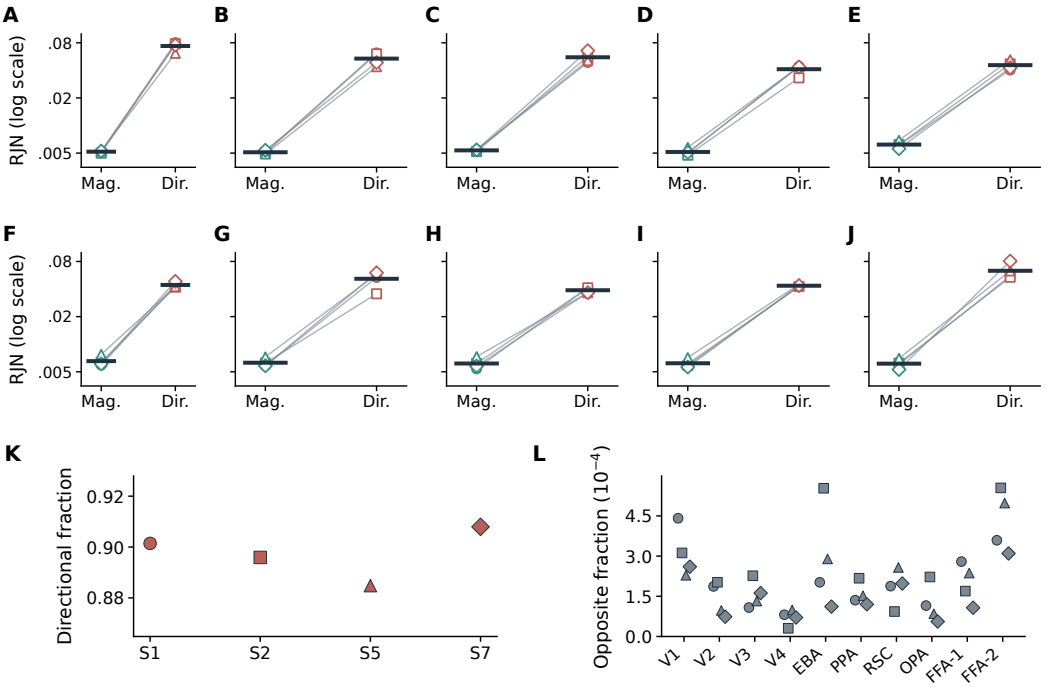

Figure S3: **Regional components across participants. A–J**, magnitude and directional RJN in V1, V2, V3, V4, EBA, PPA, RSC, OPA, FFA-1 and FFA-2. Each participant value averages 24 representation–layer conditions; lines join the two components and black segments mark four-participant means. All ten panels share logarithmic limits from 0.0035 to 0.1. **K**, directional fractions from participant means across ROIs. **L**, regional fractions of pairs with negative weighted-Jacobian inner product, in units of 10^{-4}. Circle, square, triangle and diamond identify S1, S2, S5 and S7.

RJN are 0.019742, −0.000060 and 0.019802; total RJN is higher in FFA-2 in all 96 model conditions. These conditions are repeated analyses of four participants.

We additionally compared outputs on shared support in noise ceiling (NC), predictive R^2 and log sensitivity energy $\log \max(T_C, 10^{-300})$. Seven prespecified covariate sets comprise each variable alone, each pair and all three jointly. For each set and model condition, pooled empirical quartiles define joint covariate strata; a stratum is retained when it contains at least five finite outputs from each FFA. For outcome $q \in \{\widehat{\mathcal{N}}_J, \widehat{\mathcal{N}}_{\mathrm{mag}}, \widehat{\mathcal{N}}_{\mathrm{dir}}\}$, the conditional contrast is

$$\Delta_q^{\mathrm{adj}} = \frac{\sum_{h\in\mathcal{H}} \min(n_{1h}, n_{2h})(\bar{q}_{2h} - \bar{q}_{1h})}{\sum_{h\in\mathcal{H}} \min(n_{1h}, n_{2h})},$$

where $\mathcal{H}$ is the retained set of strata and n_{rh} is the output count for FFA-r in stratum h. The raw retained contrast instead subtracts the two ROI means over all retained outputs without stratum weighting. It separates the change in available support from the conditional weighting; the original raw contrast covers all valid outputs. We average condition contrasts within participants and then participants equally. All 96 conditions are estimable for every covariate set.

With all three covariates, the equal-participant mean adjusted total contrast is 0.00480, while the mean raw contrast on its retained outputs is 0.02265 (Table S2). S7 has the narrowest joint support: FFA-1 coverage reaches 16% and FFA-2 coverage 44% in the least-covered condition. Stratification also leaves within-bin imbalance. Across the 96 joint cells, the mean absolute standardized residual differences are 0.086 for NC, 0.110 for R^2 and 0.059 for log energy (maxima 0.225, 0.215 and 0.258). Thus the joint contrast describes a smaller, approximately balanced output population, especially for S7.

The equal-participant adjusted total contrasts for NC alone, R^2 alone and log energy alone are 0.00264, 0.00803 and 0.00525; for NC+R^2, NC+log energy and R^2+log energy they are 0.00398, 0.00345 and 0.00550. S2 illustrates specification sensitivity: its NC-only mean is 0.00004 with 6/24 positive conditions, whereas the joint mean is 0.00419 with 15/24 positive conditions. Each covariate

Table S2: FFA-2 minus FFA-1 RJN under joint NC, R^2 and log-energy stratification. Entries are means over 24 representation–layer conditions per participant. Raw all uses all valid outputs; raw retained uses only outputs in shared strata; adjusted uses the stated overlap weights. Coverage is the median percentage of each ROI's eligible outputs retained across the 24 conditions. The last column counts conditions with positive adjusted total contrast. All RJN differences are unitless.

	Raw all	Raw retained	Adjusted total	Adjusted magnitude	Adjusted direction	Coverage (%) FFA-1/FFA-2	Positive total
S1	0.01267	0.01559	0.00344	−0.00009	0.00353	67/83	20/24
S2	0.01151	0.01573	0.00419	0.00020	0.00399	68/70	15/24
S5	0.01805	0.02204	0.00528	−0.00002	0.00530	73/77	24/24
S7	0.03673	0.03723	0.00629	−0.00034	0.00663	34/59	23/24

set defines its own retained output population and conditional comparison. NC estimates response reliability, R^2 summarizes predictive fit, and T_C is computed from the same Jacobian field as RJN; these comparisons characterize fitted models and do not identify a causal regional effect.

The existing fit–RJN associations show related heterogeneity at the output level. Across the 960 participant–representation–layer–ROI cells, the equal-cell mean within-cell Pearson correlation is -0.70475 before adjustment, -0.05459 after residualizing NC, and $+0.46533$ after residualizing NC and log energy. The respective positive/negative cell counts are 0/960, 403/557 and 943/17. For NC alone, FFA-1 has 44/52 positive/negative cells (mean -0.041), FFA-2 has 89/7 (mean $+0.176$), OPA has 0/96 (mean -0.193), and RSC has 94/2 (mean $+0.153$). The NC-only mean across all ten ROIs ranges from -0.093 for S1 to $+0.020$ for S7. These are means of correlations computed separately over outputs within each cell, rather than a pooled correlation over voxel–model evaluations. Including log energy changes the residual association being summarized, and the mean sign does not describe every cell.

D.3 Stimulus associations and response-conditioned dispersion

For each image and cell, we average node scores and signed predicted or recorded responses separately over the aligned outputs. Spearman correlation then compares these ROI profiles across images (Figure 4). Each participant's recorded ROI profile is reused for its twenty-four representation–layer conditions. Cross-participant score agreement averages 1,440 correlations per component: six participant pairs, two representations, twelve layers, and ten ROIs. A regional summary averages the corresponding 144 correlations, matching profiles by image identity and representation–layer condition.

For response-conditioned dispersion, images are sorted by recorded ROI response, with canonical image identity breaking ties, and divided into twenty groups of fifty. For component c, define

$$W_c = \frac{\operatorname{median}_{b=1}^{20}\{\mathrm{IQR}(d_c \mid b)\}}{\mathrm{IQR}(d_c)}.$$

The global denominator is positive in all evaluated cells. W_c compares typical within-stratum spread with overall spread; it can exceed one. The response bins collect similar ranks without requiring identical absolute responses, and the IQR ratio does not estimate residual or explained variance. Regional summaries average the 96 condition-specific ratios (Figure S4).

For the top-decile comparison, let S_c contain the 100 images with the largest component-c node scores, and let Y contain the 100 images with the largest signed recorded ROI responses. Their overlap is $|S_c \cap Y|/100$. Sets are formed within each condition and ROI, with canonical image identity breaking ties. Regional curves average the 96 condition-specific overlaps. This comparison measures agreement in selecting the upper-ranked images, complementing the all-image Spearman correlations.

D.4 Layer profiles and paired geometry comparisons

For each participant and representation, a layer profile first averages the ten ROI means, followed by an equal average over participants (Figure 6). Each layer supplies a different representation and fitted head. The directional fraction is computed from the corresponding mean directional and total scores.

Geometry comparisons form 80 strata: four participants, two representations and ten ROIs. Within each stratum, Pearson correlation compares the twelve ROI-mean R^2 values with the twelve ROI-mean RJN values. Figure 7 connects the raw-coordinate and covariance results within the same stratum. A sign change is counted when the two paired correlations have opposite signs.

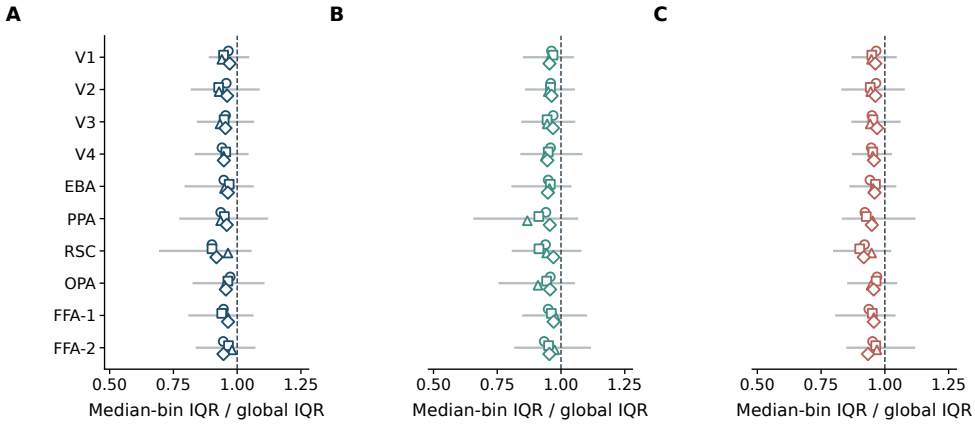

Figure S4: **Mapping-score dispersion across recorded-response strata. A**, total; **B**, magnitude; **C**, direction. The statistic W_c compares median within-bin and global score IQRs after grouping images by recorded-response rank. Circles, squares, triangles and diamonds denote S1, S2, S5 and S7 participant means, respectively; horizontal segments span the minimum to maximum W_c across model conditions for each ROI. The dashed line marks $W_c = 1$, where the median within-bin IQR equals the global IQR.

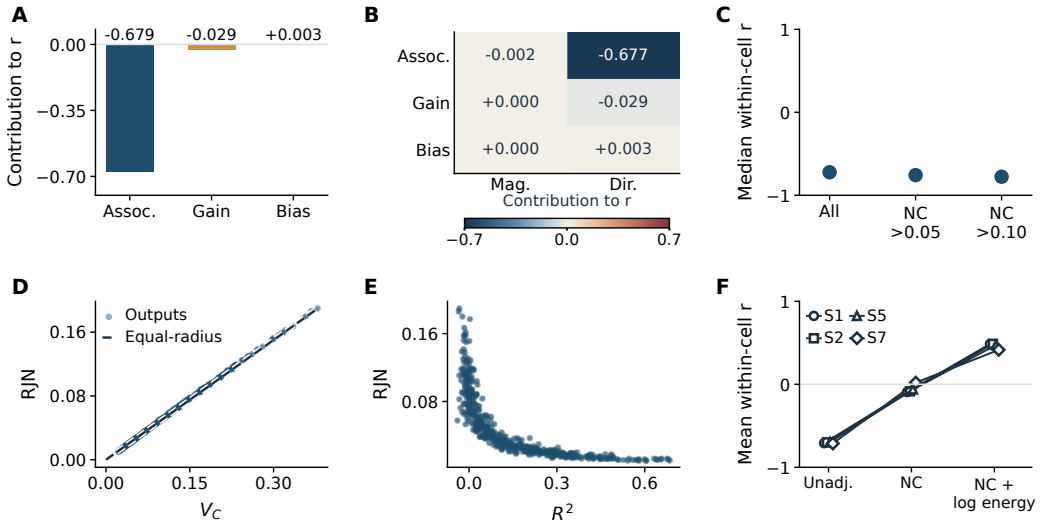

Figure S5: **Components and covariates of the fit–RJN association. A**, mean covariance contributions from squared response association, gain consistency and bias. **B**, their six contributions crossed with magnitude and direction, under the same correlation normalization. **C**, median within-cell correlations for all outputs and two noise-ceiling thresholds. **D/E**, RJN against normalized gradient-energy variation V_C and predictive R^2 for an anatomically spaced S5/CLIP/layer-10 sample. The dashed line in D is the equal-radius identity $nV_C/[2(n-1)]$. **F**, participant-average correlations before and after adjustment for noise ceiling and log sensitivity energy. Circle, square, triangle and diamond denote S1, S2, S5 and S7.

E Predictive-fit accounting and conditioning

For a voxel, let ρ be the Pearson correlation between prediction $\hat{y}$ and recorded response y, $a = \mathrm{SD}_n(\hat{y})/\mathrm{SD}_n(y)$, and $b = (\bar{\hat{y}} - \bar{y})/\mathrm{SD}_n(y)$. The moments use the same reference images and denominator n. Expanding the mean squared prediction error gives

$$R^2 = 2a\rho - a^2 - b^2 = \rho^2 - (a-\rho)^2 - b^2. \tag{13}$$

The three terms describe squared response association, a gain-consistency penalty and a mean-bias penalty. Both response and prediction standard deviations are positive for the 4,535,136 evaluated

voxel–model combinations. Among these, 651,901 have negative Pearson correlation; the squared-association term consequently needs to be interpreted with the gain penalty.

For a cell, let s_Y and s_N be the output-wise standard deviations of predictive R^2 and $\widehat{\mathcal{N}}_J$. Linearity of covariance gives three contributions

$$\frac{\mathrm{Cov}(\rho^2, \widehat{\mathcal{N}}_J)}{s_Y s_N}, \qquad -\frac{\mathrm{Cov}((a-\rho)^2, \widehat{\mathcal{N}}_J)}{s_Y s_N}, \qquad -\frac{\mathrm{Cov}(b^2, \widehat{\mathcal{N}}_J)}{s_Y s_N},$$

which sum to the within-cell correlation. Bilinearity with $\widehat{\mathcal{N}}_J = \widehat{\mathcal{N}}_{\mathrm{mag}} + \widehat{\mathcal{N}}_{\mathrm{dir}}$ gives six crossed contributions using this common denominator, with row and column sums recovering the marginal accountings. Contributions are computed within cells before equal-cell averaging. The resulting identity has maximum numerical discrepancy below 10^{-14}. A contribution's size depends on component spread and covariance under the common normalization; the accounting quantifies neither unique predictive information nor causal effects.

Partial correlations use finite voxel-level outputs shared by both target quantities and all covariates, requiring at least thirty per cell. The noise-ceiling covariate is the voxelwise NSD-style explainable-variance ceiling, represented as a unitless fraction on $[0, 1]$. Following the NSD convention, a noise-ceiling signal-to-noise ratio ncsnr for responses averaged over n_{rep} repetitions corresponds to $\mathrm{NC} = \mathrm{ncsnr}^2/(\mathrm{ncsnr}^2 + 1/n_{\mathrm{rep}})$ [2]; the analysis uses the corresponding precomputed per-voxel values rather than re-estimating reliability. Covariates with standard deviation greater than 10^{-12} are standardized. Ordinary least squares with an intercept removes the same covariate matrix from both targets; Pearson correlation of the residuals gives the partial coefficient. Residual standard deviations below 10^{-14} make it undefined. The energy covariate is $\log \max(T_C, 10^{-300})$. All 960 primary cells yield finite coefficients in the three analyses. The regressions operate on voxel-level summary statistics with the encoders unchanged. Since energy comes from the same sensitivity field, its inclusion changes the association being characterized.

E.1 Relation to normalized gradient-energy variation

Across the same 960 cells, the mean within-cell Pearson correlation between RJN and V_C is 0.99954, and the mean Spearman correlation is 0.99938. Their minima are 0.99049 and 0.98695. Mean absolute deviation from $V_C/2$ is 0.00205 after equal-cell averaging. Equation 9 relates this agreement to pair-scale weighting and its covariance with discrepancy. The scatter sample in Figure S5D–E comprises 500 outputs selected at equally spaced anatomical indices from the union of the ten S5/CLIP/layer-10 ROIs; the reported correlation summaries use the full valid cell populations.

F What normalization changes on the same models

The baseline comparisons ask where pair-relative normalization changes the description of a Jacobian field: across outputs, between regions, or over individual images. All scores use the same frozen heads, shared images and output identities. Before computing new contrasts, we specified a balanced subset comprising all four participants, both representations and layers 1, 6 and 12, spanning early, middle and final representations. Each of these 24 conditions retained all 1,000 shared images and every output in the union of the ten existing ROI masks, giving 162,504 output–condition rows. ROI averages used the original masks without a significance gate. We additionally compared the stored RJN and energy components across all 96 conditions. Condition summaries and voxel associations below describe these fitted models; participant summaries weight the four participants equally.

Separating the score formula from geometry. For the unweighted Jacobian j_i in the stored standardized feature and output coordinates, the original definitions are

$$a_i = \|j_i - \bar{j}\|_1, \qquad \bar{a} = \frac{1}{n}\sum_i a_i, \tag{14}$$

$$\mathrm{JNE} = \left\{ \frac{1}{n}\sum_i (a_i - \bar{a})^2 \right\}^{1/2}, \qquad \mathrm{JNE\text{-}SS}_i = (a_i - \bar{a})^2. \tag{15}$$

These follow Section 2.4 and Appendix A.10 of Yang et al. [28]. In particular, the image mean of JNE-SS equals JNE^2 for each output. Squaring the centered distance assigns high scores to distances on either side of $\bar{a}$. We recomputed these formulas on the current heads rather than importing scores

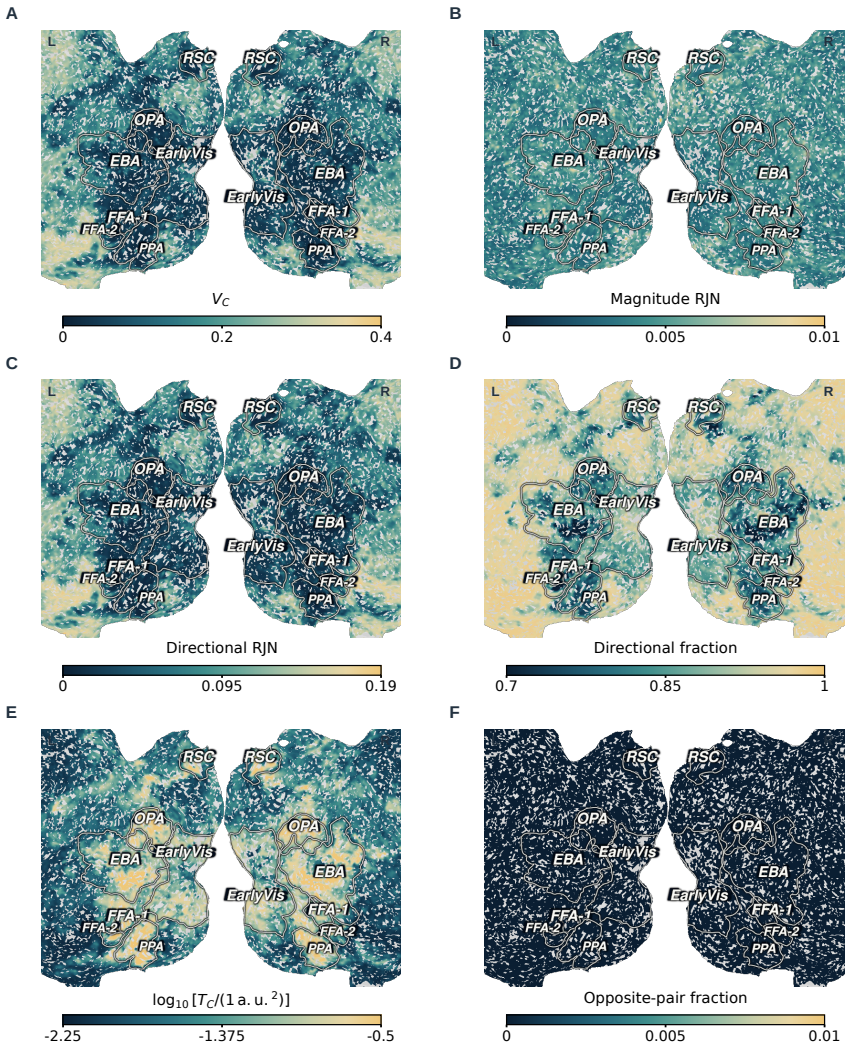

Figure S6: **Cortical fields of mapping variation and sensitivity.** All panels show S5/CLIP/layer 10. **A**, normalized gradient-energy variation V_C; **B**, magnitude RJN; **C**, directional RJN; **D**, directional fraction; **E**, $\log_{10}[T_C/(1\,\mathrm{a.u.}^2)]$; **F**, fraction of pairs with negative weighted-Jacobian inner product. The fields use the same bilateral window and native ROI contours, with scales displayed individually.

from the original study. A separate covariance-weighted analytic variant replaces j_i by $g_i = j_i C^{1/2}$ in the same formulas. This variant shares RJN's geometry but is not the original JNE; its L1 contraction remains coordinate dependent. Both versions retain their native scale.

The natural energy components and conserved nodes are those derived in Appendix B.9. They use the same g_i and C as RJN. Their finite-reference convention includes the zero diagonal, whereas RJN averages distinct pairs; multiplying the energy scores by $1000/999$ reconciles this convention without changing any correlation or selected image. Original JNE versus covariance RJN changes both formula and geometry; the weighted JNE variant isolates a formula comparison within the common geometry.

Output ordering and regional contrasts. Across the 960 ROI–condition cells in the full bank, the median within-ROI voxel Pearson correlations with the corresponding energy quantities were 0.9998 for total RJN, 0.9960 for magnitude and 0.9999 for direction (ranges 0.9905–1.0000, 0.9642–0.9995 and 0.9918–1.0000). Thus the close total-score agreement extends to both components. The pair-weighting identity in Appendix B.9 quantifies the remaining difference: the ROI-mean covariance correction divided by ROI-mean RJN had median -0.0289 (range -0.0731 to -0.0050).

Model means and within-region output scores describe different variation, as the JNE comparison illustrates. Across the 24 pooled-ROI model means, Pearson correlations of RJN with original

Table S3: **Stimulus-profile comparisons on identical heads, images and ROI outputs.** Median [minimum, maximum] across all 240 ROI–condition profiles from the balanced 24-condition subset. Each profile contains 1,000 images. Energy magnitude and direction nodes are compared with the matching RJN component; the other rows use total RJN. The weighted JNE-SS row is an analytic covariance variant. Top-100 overlap is an intersection fraction, not a Jaccard index.

Comparator	Pearson r	Spearman ρ	Top-100 overlap
Original JNE-SS	.659 [.173, .819]	.000 [−.044, .051]	.70 [.42, .93]
Weighted JNE-SS	.703 [.397, .836]	.135 [.003, .250]	.76 [.39, .90]
V_C nodes	.943 [.860, .978]	.921 [.826, .974]	.84 [.58, .95]
V_{mag} nodes	.983 [.902, .993]	.998 [.985, .999]	.87 [.73, .95]
V_{dir} nodes	.939 [.859, .968]	.924 [.828, .972]	.81 [.56, .91]

Table S4: **Participant-level FFA-2 minus FFA-1 differences in the balanced subset.** Each participant averages six conditions (two representations and three layers). Mean weights the four participants equally; the last column counts positive condition differences among 24. Values retain each metric's native scale. The JNE-SS rows average squared JNE over outputs, as required by the image-mean identity.

Metric	S1	S2	S5	S7	Mean	Positive
RJN	.01302	.01475	.01863	.03709	.02087	24/24
RJN magnitude	−.00016	.00013	−.00031	−.00004	−.00010	10/24
RJN direction	.01318	.01461	.01894	.03713	.02097	24/24
V_C	.02548	.02961	.03706	.07296	.04128	24/24
V_{mag}	−.00023	.00041	−.00042	.00020	−.00001	11/24
V_{dir}	.02571	.02919	.03748	.07277	.04129	24/24
Original JNE	−.00225	−.00292	−.01043	−.03528	−.01272	2/24
Weighted JNE	−.01855	.00299	−.02443	−.09045	−.03261	3/24
Mean original JNE-SS	.00064	.00117	−.00569	−.02386	−.00693	5/24
Mean weighted JNE-SS	−.00857	.01745	−.00575	−.05791	−.01370	8/24

and weighted JNE were 0.923 and 0.863, respectively. Within the 240 ROI–condition cells, their median voxel correlations were instead -0.737 and -0.682. JNE retains the absolute Jacobian scale, whereas RJN compares differences relative to each pair's scale. Agreement between model means can therefore coexist with opposite ordering of outputs within a region.

Table S4 reports the paired FFA contrasts in each score's native units. Total RJN, V_C and their directional components were higher in FFA-2 in all 24 selected conditions, with positive participant-mean differences for all four participants. Original JNE was higher in FFA-2 in only 2 of 24 conditions and its participant-mean differences were all negative. The weighted JNE variant was positive in 3 of 24 conditions. The energy comparison therefore reproduces the regional ordering, while the distance-dispersion summaries describe a different regional pattern.

Allocation over individual images. For each condition and ROI, we averaged each stimulus score over exactly the same outputs. Table S3 summarizes all 240 resulting profiles; each correlation uses all 1,000 images. The top-decile overlap is the intersection of the two top-100 sets divided by 100, with canonical image ID breaking exact ties. The comparison uses the full image profiles rather than an image selection based on disagreement. Natural energy nodes showed substantial agreement with RJN: median rank correlations were 0.921 for total and 0.924 for direction, and 0.998 for magnitude. Nevertheless, the total-node top-decile overlap was 0.84, rather than identical image selection. Original and weighted JNE-SS gave much lower rank correlations, despite appreciable Pearson correlations and top-decile overlap; their squared centered-distance construction differs from both pair-based allocations.

The distinction is clearest at the stimulus level. Energy components and RJN components organize outputs almost identically in this bank, while pair-relative weighting changes some stimulus rankings, particularly for total and directional scores. Decomposition and mean-conserving allocation are shared capabilities of RJN and the energy baseline; the observed differences concern the weighting and organization of their scores.

The comparison preserves the original bank's checkpoint hashes, canonical image IDs, output IDs and covariance hashes. Recomputed RJN ROI profiles differed from their stored counterparts by at most 5.12×10^{-10}; the maximum discrepancy among compared scalar fields was 4.34×10^{-9}.

Direct small-kernel checks verified the component and node identities. All comparisons use the frozen parameters.

G How geometry estimation and fitting change the description

A mapping description can change because the feature sample changes its geometry or because fitting yields a different function. We examine these sources separately, with condition and output selections fixed before computing their effects. Both analyses retain the original image identities and ROI definitions; numerical evaluation is checked alongside them.

G.1 Reestimating covariance with the head fixed

For each of the 96 original heads, we selected the first, middle and last anatomically ordered output in each ROI. Scores use all 1,000 reference images and all distinct pairs. Twenty bootstrap samples of 1,000 feature rows, drawn with replacement, each define a new centered empirical covariance with denominator 1,000. The encoder and evaluation images remain fixed. We use the same twenty index patterns across conditions, generated with seed 20260924. Each covariance is evaluated through its positive-semidefinite square root without diagonal regularization. Repeated feature rows can reduce its rank; its null directions receive zero weight.

The comparison target is the score computed with the full-reference covariance on these same selected outputs. Within each ROI, we average its three output scores and compare each reestimated-geometry value with this target. Layerwise correlations compare the twelve selected-output ROI means of RJN with their fixed held-out fit means, within each participant, representation and ROI. The twenty paired resamples describe empirical sensitivity of this calculation to the feature sample used for geometry estimation. Their observed spread characterizes geometry estimation within the existing participants and feature samples, rather than independent population replications or new participants.

G.2 Alternative fits with a common data split

The refit analysis covers layer 10 of both representations for all four participants, selected because the existing stimulus examples use this layer. We fit two new initializations, seeds 1 and 2, for each of these eight conditions. Each retains the original architecture, full trained output set and fixed training/validation/reference split. Training uses the same mean squared objective, AdamW settings, maximum epoch count, validation selection and patience specified in Appendix C.1. Accumulating gradients over batches of 512 images assembles one full-objective update per epoch; this changes the floating-point summation order from the historical single forward pass. The historical original head supplies the third fit, alongside the two newly fitted initializations; it is not a newly reproduced seed-0 head.

Evaluation uses the same covariance and 1,000 reference images for all three heads in a condition. Before computing scores, we selected 32 evenly spaced anatomical output indices in each ROI; the selected identities are held fixed across heads. We compare voxel-score ranks within these subsets and ROI-mean stimulus-score ranks over the 1,000 images. Each condition contributes three pairs of heads. Thus each component has 240 ROI–condition–head-pair profile correlations, of which 160 compare an original head with a new fit and 80 compare the two new fits. FFA contrasts use the same 32-output subsets. Their scope is this layer and these output subsets.

G.3 Variation in scores, image orderings and regional contrasts

Across the 19,200 ROI–condition–resample comparisons per component, median absolute relative changes were 1.58% for total RJN, 1.93% for magnitude and 1.78% for direction; their maxima were 9.90%, 18.01% and 12.15%. The median absolute changes in the corresponding layerwise fit correlations were 0.011, 0.014 and 0.013. Total and directional correlations retained their signs in all 1,600 paired comparisons per component; magnitude changed sign in 28. All correlations were defined. These summaries compare reestimated covariance with full-reference covariance, within the fixed three-output ROI subsets (Figure S7A–B).

Refitting preserved image ordering more strongly for total and directional scores than for magnitude (Table S10; Figure S7C). Magnitude voxel rankings varied substantially, including negative correlations in some 32-output subsets. The original–new and new–new profile comparisons gave similar medians: total 0.931 and 0.931, magnitude 0.664 and 0.664, and direction 0.945 and 0.943, respectively. Thus an image-level regional average can remain more consistent across fitted heads than individual magnitude scores within that region.

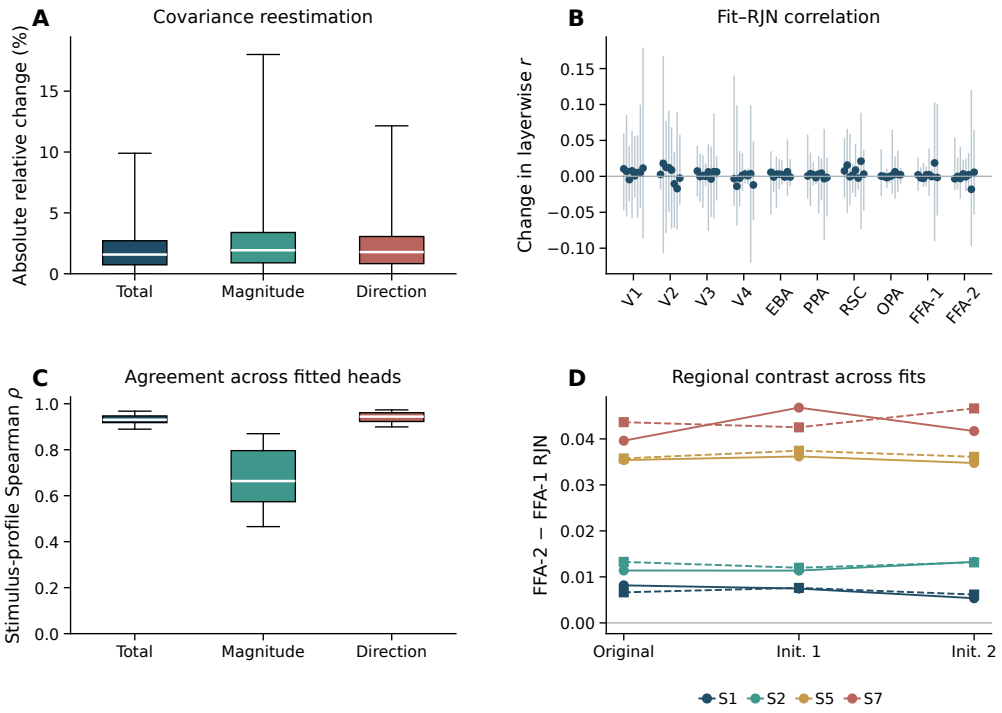

Figure S7: **Geometry estimation and variation between fitted heads. A**, absolute relative changes in ROI-mean scores after reestimating covariance; each component pools 96 conditions, ten ROIs and twenty resamples. **B**, changes in total-RJN versus fit correlations across layers. Dots are medians over twenty resamples and lines span their observed range; each ROI has eight participant–representation strata. **C**, ROI stimulus-profile rank agreement across pairs of fitted heads at layer 10. Boxes in A and C show the interquartile range and median; whiskers span all observed values. **D**, FFA contrasts within each fitted condition. Color identifies participant; solid lines and circles denote CLIP, dashed lines and squares supervised ViT. Geometry summaries use three fixed outputs per ROI; refit summaries use 32.

The total and directional FFA-2 minus FFA-1 contrasts remained positive for all 24 condition–head combinations; magnitude was positive in 15. Averaging the two representations within participant, total differences for the original and two new heads were 0.00740, 0.00754, 0.00576 in S1; 0.01232, 0.01168, 0.01321 in S2; 0.03557, 0.03681, 0.03544 in S5; and 0.04162, 0.04466, 0.04416 in S7 (Figure S7D). Across all trained outputs, each new head's held-out mean squared error differed from its original by between -0.092% and $+0.120\%$. Similar prediction error therefore accompanies different degrees of component-level agreement.

The selected original-head scores reproduced the bank to a maximum absolute difference of 1.21×10^{-8} and its held-out R^2 to 6.60×10^{-8}. Across the new analyses, analytic derivatives differed from automatic differentiation by at most 4.29×10^{-8}; the direct pair-kernel discrepancy was below 7.34×10^{-7}. Component conservation held to 5.56×10^{-17}. These checks establish the numerical agreement of the evaluations used for the geometry and refit comparisons.

H Geometry from a separate feature sample

To separate the feature sample used for geometry from the images at which mappings are compared, we estimated covariance from the original 8,000 training features in each condition. The reference and validation images are excluded, using the unchanged split in Appendix C.1. The covariance uses its own training-feature mean and denominator 8,000. We compare it with the reference-feature covariance while holding the fitted function, 1,000 evaluation images and output identities fixed. This changes the sample defining the perturbation geometry, with all neural measurements drawn from the same NSD study.

Table S5: Layerwise fit–RJN associations under five fixed geometries. Each coefficient uses twelve layers and the same three outputs per ROI; summaries cover 80 participant–representation–ROI strata. Entries give median [minimum, maximum]. Sign counts are negative/zero/positive.

Geometry	Pearson	Signs	Spearman	Signs
Reference covariance	-0.707 $[-0.989, 0.553]$	75/0/5	-0.531 $[-0.923, 0.650]$	75/0/5
Training covariance	-0.692 $[-0.990, 0.558]$	74/0/6	-0.503 $[-0.937, 0.692]$	74/1/5
$\lambda = 0.1$	-0.726 $[-0.988, 0.522]$	75/0/5	-0.528 $[-0.895, 0.434]$	75/0/5
$\lambda = 0.5$	-0.689 $[-0.981, 0.394]$	77/0/3	-0.490 $[-0.909, 0.427]$	71/0/9
$\lambda = 1$	-0.134 $[-0.900, 0.755]$	47/0/33	-0.087 $[-0.923, 0.839]$	43/2/35

Table S6: Component associations on the same selected outputs. Entries are median coefficients; sign-count conventions match Table S5.

Component	Geometry	Pearson	Signs	Spearman	Signs
Magnitude	Reference covariance	-0.361	65/0/15	-0.262	52/1/27
Magnitude	Training covariance	-0.380	63/0/17	-0.252	55/0/25
Magnitude	$\lambda = 0.1$	-0.355	63/0/17	-0.227	55/0/25
Magnitude	$\lambda = 0.5$	-0.353	62/0/18	-0.196	55/1/24
Magnitude	$\lambda = 1$	-0.397	64/0/16	-0.234	54/0/26
Direction	Reference covariance	-0.707	75/0/5	-0.545	75/0/5
Direction	Training covariance	-0.704	75/0/5	-0.580	72/0/8
Direction	$\lambda = 0.1$	-0.741	75/0/5	-0.584	75/0/5
Direction	$\lambda = 0.5$	-0.677	75/0/5	-0.517	67/1/12
Direction	$\lambda = 1$	-0.033	41/0/39	0.024	40/0/40

We also evaluate the fixed family

$$C_\lambda = (1-\lambda)C_{\text{train}} + \lambda \frac{\text{tr}(C_{\text{train}})}{768} I, \qquad \lambda \in \{0, 0.1, 0.5, 1\}.$$

The eigenvectors of the training covariance are retained, and each eigenvalue is interpolated toward their mean. The positive trace makes the last endpoint proportional to identity, giving the same RJN as unscaled identity. Intermediate values progressively weight coordinate directions more equally. These specified values describe a perturbation family; no fit association is used to select a preferred member. The original reference covariance and four members of the training-based family give five comparisons per condition.

All 96 conditions use the same three anatomically ordered outputs per ROI as the covariance-resampling analysis (Appendix G). ROI means average these outputs, and each of the 80 participant–representation–ROI strata contributes twelve layer values. We compute both Pearson and Spearman correlations between the selected-output means of predictive fit and each RJN component. Pearson keeps the coefficient used in the main geometry analysis; Spearman describes its rank counterpart. The output subsets are smaller than those underlying Figure 7. Comparisons between geometries use identical subsets, and the accompanying tables summarize both coefficients and all sign categories.

H.1 Observed associations and regional contrasts

With reference covariance, the selected-output means give 75 negative and five positive total-RJN Pearson associations across layers (Table S5). Estimating covariance from the disjoint training images preserves 79 of these 80 signs; the median absolute coefficient change is 0.024, with maximum 0.116. Increasing the isotropic contribution changes the layerwise relationships more substantially. At identity, 33 Pearson coefficients are positive and 47 are negative, with 38 strict sign reversals relative to reference covariance. Across the training-based family, the median coefficient is not monotonic in λ.

The rank analysis gives the same broad dependence on geometry while retaining differences in individual signs. Training covariance yields 74 negative, one zero and five positive Spearman coefficients; identity yields 43 negative, two zero and 35 positive values. Three and forty strata, respectively, change sign category relative to reference covariance, including transitions to zero. Tables S5 and S6 show total and component summaries. All coefficients are defined in these comparisons.

The sparse output selection also changes regional averages. Under reference covariance, the FFA total contrast is positive in 49 of 96 conditions on these subsets, compared with all 96 when averaging

Table S7: FFA contrasts on the geometry-analysis subsets. Differences are FFA-2 minus FFA-1, summarized over all 96 conditions with equal weight. The final columns count strictly positive contrasts for total, magnitude and direction.

Geometry	Mean total	Total range	Total	Mag.	Dir.
Reference covariance	0.0260	$[-0.0609, 0.0993]$	49/96	34/96	48/96
Training covariance	0.0257	$[-0.0589, 0.0992]$	49/96	42/96	49/96
$\lambda = 0.1$	0.0206	$[-0.0439, 0.0731]$	48/96	42/96	48/96
$\lambda = 0.5$	0.0118	$[-0.0295, 0.0566]$	48/96	53/96	48/96
$\lambda = 1$	0.0050	$[-0.0280, 0.0403]$	52/96	61/96	53/96

Table S8: Low-sensitivity exclusion changes the regional contrast. All 96 original conditions are included. Differences are dimensionless FFA-2 minus FFA-1 scores, averaged with equal participant weight. The last column gives the number of positive condition contrasts, which is identical for total RJN and V_C.

Excluded percentile	Total RJN difference	V_C difference	Positive / 96
0	0.019742	0.038867	96
1	0.017384	0.034164	96
5	0.012965	0.025415	96
10	0.009190	0.017993	95
20	0.005956	0.011651	92

the original full ROIs. Training covariance retains 49 positive contrasts; identity gives 52 (Table S7). Positive average differences therefore coexist with heterogeneous condition signs. These anatomically fixed subsets allow geometry to be changed at constant output identities, while the main regional result concerns the larger original output populations. Recomputed reference-covariance scores agree with the stored scores for the selected outputs to a maximum absolute difference of 1.21×10^{-8}.

I Relative variation when sensitivity is weak

Weak absolute sensitivity can accompany substantial relative mapping variation. To interpret that combination, we ask three distinct questions: how the regional average changes when low-energy outputs are excluded, how the score changes when pair normalization receives an absolute floor, and how the gradients themselves change between fitted heads. Filtering changes the output population, flooring changes the diagnostic, and refitting changes the function being described. The analyses use four participants and report descriptive comparisons within their fitted models.

I.1 Excluding low-sensitivity outputs in all 96 conditions

For each original condition, we formed the T_C distribution across the unique union of the ten ROIs and computed its 0, 1, 5, 10 and 20 percentiles. The zero cutoff retains all outputs; other cutoffs retain outputs at or above the condition-specific threshold. RJN and V_C, including their components, always use identical retained outputs. Each regional score is the mean over that ROI's retained outputs. Its rank comparison uses the ten regional means before and after exclusion. FFA contrasts first average the 24 conditions within each participant and then give the four participants equal weight.

At the 20th-percentile cutoff, all 960 ROI–condition cells remain populated, retaining between 51.05% and 99.19% of their original outputs (median 87.27%). A pooled cutoff therefore removes different fractions within different regions. The total-RJN FFA contrast remains positive after averaging within each participant, with differences 0.001524, 0.006144, 0.004365 and 0.011789 for S1, S2, S5 and S7. Four individual conditions reverse sign: S1 CLIP layers 1–3 and S1 supervised ViT layer 2. At the 10th-percentile cutoff, only S1 CLIP layer 2 reverses. Directional contrasts have the same positive-condition counts as total RJN; magnitude has 47, 50, 55, 54 and 54 positive conditions across the five cutoffs.

At the largest cutoff, the median [minimum, maximum] Spearman correlation of the ten regional means with their unfiltered ordering is 0.903 $[0.479, 0.988]$ for total RJN and, for V_C, 0.903 $[0.442, 0.988]$. The smaller FFA contrast and occasional reversals demonstrate dependence on the selected output population. Both weightings show the same sign pattern and attenuation; their raw numerical differences use different metric scales.

Table S9: Floor diagnostics on the eight original layer-10 heads. Activation, rank and overlap columns give medians over 80 ROI–condition cells for total RJN. FFA counts refer to the eight condition contrasts. The original grid ends at 0.1; 0.3 and 1.0 follow the documented amendment.

a	Active pairs (%)	Voxel ρ	Top-100 overlap	Positive FFA
0	0.000	1.000	1.000	8/8
0.01	0.000	1.000	1.000	8/8
0.03	0.000	1.000	1.000	8/8
0.1	0.000	1.000	1.000	8/8
0.3	0.000	1.000	1.000	8/8
1	47.174	0.595	0.900	2/8

I.2 An absolute sensitivity reference in the pair denominator

We next use layer 10 of both representations for all four participants, with the same 32 anatomically selected outputs per ROI and 1,000 reference images as Appendix G. For each condition, let s_C be the median of $\sqrt{T_C}$ across the full original ten-ROI union, and fix $\tau = as_C$ for $a \in \{0, 0.01, 0.03, 0.1\}$. We replace only the pair denominator by

$$d_\tau(r, s) = \max\{r + s, 2\tau\}^2,$$

where r and s are the covariance-weighted Jacobian norms. The original total, magnitude and directional numerators, distinct-pair average and zero-pair convention are retained. This diagnostic introduces a shared sensitivity reference and changes the estimand. At fixed $\tau > 0$, its kernel satisfies $k_\tau(cg, ch) = k_{\tau/|c|}(g, h)$ for $c \neq 0$, so rescaling an individual output can change its value. Simultaneously scaling every gradient and the floor by the same positive factor preserves the diagnostic. Appendix B.2 gives deterministic perturbation bounds for a fixed floor. We compare voxel ranks, regional stimulus profiles, top-100 image sets and FFA contrasts with the zero-floor values. Pair activation records how often $r + s < 2\tau$; a floor that never activates supplies no evidence about statistical noise suppression.

After inspecting the first completed condition (S1 CLIP), in which none of the original nonzero floors activated a pair, we recorded an explicit protocol amendment adding $a \in \{0.3, 1.0\}$ for all eight original heads before computing these extension results. That initial observation informed the extension, which is reported separately from the original grid. The stronger scales test the effect of an absolute sensitivity reference; they supply neither a calibrated noise estimate nor a default threshold.

The original nonzero floor grid activates pairs in 0 of the 2,560 ROI–output–condition entries at its largest scale. Table S9 reports the median activated-pair fraction, voxel-rank agreement and top-100 overlap over the 80 ROI–condition cells; the two amended scales are 0.3 and 1.0.

At scale 0.3, 25 ROI–output–condition entries activate pairs, the largest ROI–condition mean activation is 0.314%, and the minimum top-100 overlap is 0.990. At $a = 1$, total-RJN voxel-rank agreement is 0.595 $[-0.530, 0.985]$, stimulus-profile agreement is 0.964 $[0.913, 0.993]$, and top-100 overlap is 0.900 $[0.760, 0.980]$. The FFA contrasts range from -0.002979 to 0.005404. Magnitude and direction voxel-rank medians are 0.483 and 0.677. Six of the eight FFA condition contrasts reverse sign. These changes describe the effect of imposing a common absolute sensitivity reference on the selected outputs.

I.3 Gradient drift between fitted heads and comparison with energy nodes

For the same eight conditions, we evaluate the original head and the two existing alternative fits from Appendix G, with common covariance, images and selected outputs. This analysis reuses those fitted heads. We recompute conserved total, magnitude and directional energy nodes on the identical outputs, retaining the n denominator for energy nodes and the $n-1$ distinct-pair denominator for RJN. Each component has 240 ROI–condition–head-pair comparisons of voxel ranks and 1,000-image regional profiles. Top-100 sets are compared by intersection size divided by 100; ties use ascending canonical image identity.

Original T_C quintile boundaries are computed from the full ten-ROI union in each condition, then applied to the fixed anatomical subsets. Each quintile has 16 original–alternative-head comparisons; bin identities never change between heads. For every selected output we also record

$$D = \sqrt{\frac{n^{-1}\sum_i \|g_i^{\text{new}} - g_i^{\text{original}}\|_2^2}{T_C^{\text{original}}}},$$

Table S10: Agreement across the three fitted heads, on the same fixed 32 outputs per ROI. Entries are medians [minimum, maximum] over 240 ROI–condition–head-pair comparisons per metric. Profile ranks and top-100 overlap concern the same 1,000 images.

Metric	Voxel ρ	Profile ρ	Top-100 overlap
RJN total	$0.972\ [0.908, 0.992]$	$0.931\ [0.890, 0.968]$	$0.760\ [0.640, 0.870]$
RJN magnitude	$0.355\ [-0.279, 0.780]$	$0.664\ [0.466, 0.870]$	$0.620\ [0.400, 0.770]$
RJN direction	$0.977\ [0.907, 0.995]$	$0.944\ [0.899, 0.973]$	$0.760\ [0.620, 0.890]$
V_C total	$0.974\ [0.907, 0.992]$	$0.898\ [0.830, 0.925]$	$0.760\ [0.610, 0.870]$
V_C magnitude	$0.359\ [-0.288, 0.770]$	$0.661\ [0.456, 0.871]$	$0.600\ [0.410, 0.750]$
V_C direction	$0.978\ [0.903, 0.994]$	$0.912\ [0.882, 0.945]$	$0.740\ [0.660, 0.840]$

Table S11: Refit sensitivity stratified by original sensitivity energy. Quintile 1 has the lowest original energy. Rank entries summarize 16 original–alternative-head comparisons per quintile; drift summarizes the corresponding 16 within-bin median drift ratios. All five bins use fixed original output identities.

T_C quintile	Total RJN ρ	Total V_C ρ	Relative drift D
1	$0.747\ [0.551, 0.895]$	$0.750\ [0.543, 0.895]$	$1.148\ [1.024, 1.231]$
2	$0.712\ [0.592, 0.851]$	$0.717\ [0.592, 0.854]$	$0.861\ [0.680, 0.987]$
3	$0.772\ [0.649, 0.862]$	$0.782\ [0.655, 0.877]$	$0.647\ [0.508, 0.763]$
4	$0.754\ [0.598, 0.857]$	$0.765\ [0.617, 0.867]$	$0.496\ [0.407, 0.573]$
5	$0.861\ [0.580, 0.926]$	$0.873\ [0.623, 0.929]$	$0.376\ [0.316, 0.437]$

where both Jacobians use the same covariance geometry. This quantity measures relative gradient drift induced by alternative initialization and optimization on the same data. The comparison measures agreement among fitted mappings, with biological-noise calibration and the true Jacobian unobserved. All quintiles, empty subsets and undefined correlations are retained in the result files. An independent target would be required to compare the scientific accuracy of the weightings.

I.4 Agreement of output ranks and stimulus profiles

Within-bin ranks restrict the score range, so we also report paired score changes. Median mean absolute changes in the lowest versus highest energy quintile are 0.01277 versus 0.00132 for total RJN and 0.02537 versus 0.00254 for total energy variation. Both metrics show larger refit changes in weak-sensitivity outputs, with numerical magnitudes expressed on their respective scales.

Within the lowest-energy quintile, magnitude voxel-rank agreement is $0.405\ [-0.084, 0.681]$ for RJN and $0.383\ [-0.082, 0.681]$ for the energy baseline. The lowest-quintile regional-profile analysis has 146 defined comparisons out of 160 planned ROI–condition–original/alternative pairs. On these available subsets, total-RJN profile agreement is $0.862\ [0.527, 0.963]$, compared with $0.790\ [0.414, 0.906]$ for the energy nodes. Empty subsets remain explicitly recorded.

Across all 240 paired regional comparisons, the within-comparison difference in total stimulus-profile correlation (RJN minus energy) is $0.035\ [0.008, 0.070]$; the corresponding top-100 overlap difference is $0.000\ [-0.090, 0.110]$. For the 80 original-head regional profiles, RJN versus energy-node rank agreement is $0.926\ [0.890, 0.951]$, with top-100 overlap $0.830\ [0.690, 0.950]$. The two weightings thus share sensitivity to weak outputs while allocating variation differently over images within the same fitted-model setting.

J Figure construction and reproducibility

J.1 Stimulus selection and image displays

In S5/CLIP/L10, OPA and PPA images are eligible when their recorded target-ROI response falls within 0.10 standard deviations of its median, yielding 85 OPA and 80 PPA images. The bottom and top quartiles of eligible directional scores define low and high groups. We match these groups one-to-one by minimum absolute recorded-response difference. For each ROI, three unused matches are selected in sequence by proximity of their mean response to the one-third, two-thirds and median quantiles of the eligible response band. Canonical NSD identity resolves ties.

In the low/high order of pair panels A–F in Figure 5, the twelve one-based NSD image numbers are 44706, 25320, 23994, 54914, 28342, 25571, 7040, 9866, 14444, 60979, 40841, and 67364. The six recorded-response differences are 0.00011, 0.00033, 0.00100, 0.00328, 0.00542, and 0.00238

Table S12: All pairs satisfying both response tolerances of 0.01. Image numbers are ordered low/high by directional score. Differences are absolute, except for direction, which is high minus low. Several rows share images.

ROI	Low image	High image	Recorded	Predicted	Direction
OPA	61874	42008	0.00297	0.00346	0.00712
OPA	60520	42008	0.00066	0.00703	0.00952
OPA	21990	67830	0.00429	0.00577	0.01012
OPA	21990	56419	0.00503	0.00536	0.00781
OPA	28342	25320	0.00888	0.00197	0.00700
PPA	63945	9866	0.00498	0.00716	0.01815
PPA	72171	71230	0.00757	0.00599	0.00732
PPA	46003	9866	0.00402	0.00869	0.01623
PPA	70765	52329	0.00252	0.00841	0.01504
PPA	29920	55670	0.00592	0.00049	0.00913
PPA	29920	48803	0.00151	0.00892	0.00975
PPA	30408	43985	0.00843	0.00155	0.00970

standardized units. Directional percentiles in A–F use the midrank $100\{N_<(d_i) + \frac{1}{2}N_=(d_i)\}/1000$ within each ROI's 1,000-image S5/CLIP/layer-10 reference distribution. Here $N_<(d_i)$ and $N_=(d_i)$ count scores below or equal to d_i.

J.2 Joint matching of recorded and predicted responses

We retained the original S5/CLIP/layer-10 eligibility bands: recorded ROI responses within 0.10 times their across-image standard deviation of the median. The resulting 85 OPA and 80 PPA images give 22 and 20 images in each directional-score extreme quartile. Every low–high pair is considered, giving 484 OPA and 400 PPA candidates. The primary tolerance requires both the absolute recorded-response difference and the absolute predicted-response difference to be at most 0.01 in the stored standardized response units. The responses themselves are not rescaled within these bands. Table S12 lists all qualifying pairs, with one-based NSD image numbers.

Five OPA and seven PPA pairs qualify. Maximum bipartite matching gives three and five non-overlapping pairs, respectively. Directional separation follows from the extreme-quartile selection; the additional finding is that some of these images also satisfy the two response tolerances. This search describes the available candidates in two ROIs of one fitted condition.

Using common recorded/predicted tolerances of 0.005, 0.01, 0.025 and 0.05 yields 1, 5, 24 and 71 accepted OPA pairs, and 0, 7, 32 and 61 PPA pairs. The corresponding maximum non-overlapping counts are 1, 3, 11 and 17 for OPA, and 0, 5, 13 and 17 for PPA. At the primary tolerance, matching only the recorded response would admit 137 OPA and 121 PPA pairs. Joint matching therefore narrows the candidate set substantially. These additional pairs supplement the recorded-response matches displayed in Figure 5.

J.3 Joint response matching across all fitted conditions

We extended the original OPA/PPA search in Appendix J.2 to all 96 participant–representation–layer conditions, retaining the same full-trained-output directional stimulus profile, candidate band, extreme-quartile rule, and response units. In each ROI-condition cell, images were eligible when their recorded response lay within 0.1 times its across-image standard deviation of the across-image median. We enumerated every low–high directional-quartile pair. Both the recorded and fitted-model predicted ROI response differences had to be at most the common tolerance. All 1,000 reference images were available in each cell.

At the primary 0.01 tolerance, 908 of 84,816 candidate pairs qualified (1.071%); 187 of 192 cells contained at least one pair. The median count was four pairs per cell (range 0–14). The median selected high-minus-low directional difference was 0.01613 (interquartile range 0.01190–0.02375; range 0.00392–0.14798). The directional difference is positive by construction of the quartiles, and its distribution is conditional on both selection and response matching. The sums of maximum image-disjoint bipartite matches were 346 for OPA and 363 for PPA, across separately matched cells; images, model outputs and participants can recur across cells. No inferential test treats pairs or cells as independent observations.

At 0.01, OPA contributed 444 of 43,416 candidates and PPA 464 of 41,400. By participant, accepted counts were 241, 208, 175 and 284 for S1, S2, S5 and S7, respectively, each over both representations,

Table S13: Expanded joint matching. Counts pool 96 conditions per ROI or 192 cells per tolerance. Candidate rates use every low–high extreme-quartile pair as denominator. Disjoint counts sum separate within-cell maximum bipartite matchings and do not describe a globally independent sample.

Tolerance	Candidates	Recorded only	Joint	Joint rate	Disjoint sum
0.005	84,816	12,254	220	0.259%	205
0.01	84,816	24,187	908	1.071%	709
0.025	84,816	54,211	4,922	5.803%	2,013
0.05	84,816	80,997	14,581	17.191%	2,843

12 layers and two ROIs. The original S5/CLIP/layer-10 result was reproduced exactly: OPA had 85 eligible images, five accepted pairs and three disjoint matches; PPA had 80, seven and five. Complete cell-level counts and all selected canonical image identifiers are provided in the accompanying derived tables. The interpretation concerns fitted encoding mappings under the stated geometry, output population and matching rule.

J.4 Cortical projection and display scales

Scalar fields use S5's native anatomy and a shared functional-to-anatomical transform. Cortical surfaces, ROI contours and labels come from the participant's pycortex reconstruction and overlay [9]. EarlyVis groups V1–V4 only in displays. Inflated views share an orthographic scale. Three surface adjacency-averaging steps combine 0.58 of each supported vertex's value with 0.42 times its observed-neighbor mean, using triangles whose vertices all have support. Flatmaps at 700-pixel height are Gaussian-smoothed separately within each connected supported region, with standard deviation 1.55 pixels. Dividing by the identically filtered support mask corrects boundary attenuation; regions smaller than six pixels remain unchanged. Gray marks missing cortical support; extracortical pixels are transparent. Quantitative analyses use unsmoothed values.

Color limits are robust field quantiles rounded to readable endpoints: $[0, 0.19]$ for total and directional RJN; $[0, 0.4]$ for V_C; $[0, 0.01]$ for magnitude RJN and negative-inner-product pair fraction; $[0.7, 1]$ for directional fraction; and $[-2.25, -0.5]$ for $\log_{10}[T_C/(1\,\text{a.u.}^2)]$. At most 0.8% of finite projected values exceed the complementary-field limits. Measured responses use $[-1.75, 1.75]$ standardized units.