pRF-GUIDE: MEASURED RECEPTIVE FIELDS GUIDE fMRI-TO-IMAGE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learned neural readouts can infer spatial assignments from fMRI, leaving unclear whether independently measured retinotopy contributes additional information. We introduce pRF-Guide, which reweights response-dependent attention with quality-weighted population receptive-field footprints. Matched training controls separate the effect of a structured prior from that of its correct voxel correspondence. Across four NSD participants and three training seeds, true geometry improves reconstructed-image spatial cosine by 0.0039 over a retrained shuffled prior, with favorable participant means throughout. A separately refitted single-participant study supports the contribution of joint footprint–scale correspondence, while fixed-weight interventions on pRF centers and widths further link measured geometry to reconstruction outcomes. Measured retinotopic correspondence contributes information beyond learned spatial capacity, but does not uniformly improve every reconstruction component or metric.

1 INTRODUCTION

Generative models reconstruct recognizable natural images from functional magnetic resonance imaging (fMRI) by using decoded visual features to condition a strong image prior (Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023; Scotti et al., 2023). Recognition leaves spatial organization underdetermined: an image can preserve a scene category while moving an object or changing its extent. A neural interface must assign responses to visual locations as well as recover semantic content.

Existing systems learn this assignment in different ways. MindEye and MindEye2 align neural predictions with CLIP and combine semantic and low-level conditioning (Scotti et al., 2023; 2024). Brain-IT uses interactions among functionally organized brain tokens to predict localized features (Beliy et al., 2026). NeuroPictor uses multi-individual pretraining and multilevel modulation, while SynMind addresses semantic hallucination (Huo et al., 2024; Yang et al., 2026). These systems motivate a specific question: how much does independently measured visual-field correspondence add to a readout that can already learn spatial assignments?

Population receptive fields (pRFs) characterize the visual-field position and extent to which a population of neurons within a cortical voxel responds (Dumoulin & Wandell, 2008; Kay et al., 2013). These estimates supply voxel-specific spatial metadata for fMRI decoding. NSD pairs natural-image responses with retinotopic measurements that support this test (Allen et al., 2022). Because a pRF is a distributed sensitivity profile with variable fit quality, we use it as a soft preference while retaining response-dependent assignment.

We introduce pRF-Guide, which integrates measured pRF footprints over a visual grid and applies a quality-weighted relative log-mass bias to attention. True, shuffled, and absent geometry share trainable capacity. Their comparison separates the benefit of a structured prior from the additional benefit of its correct voxel correspondence. In a four-participant, three-seed study, the latter improves reconstructed-image spatial features in every participant mean, with smaller effects than the overall gain over geometry-free attention.

A separately refitted Participant 1 study tests this contribution through complementary controls. Joint footprint–scale shuffling retrains both neural branches; center and width interventions retain learned weights and routing scales. Correct joint correspondence improves spatial features and regional RGB

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.

Table 1: Spatial-readout mechanisms and their information sources.

Method / input	Source of spatial assignment
IRFA (Le et al., 2025)	
Macaque multi-unit activity	Learned spatial and feature-selective attention.
Brain-IT (Beliy et al., 2026)	
Human fMRI	Functional voxel-cluster tokens interact to predict localized image features.
pRF-Guide (ours)	
Human fMRI	Measured pRF footprints bias content-dependent attention; retrained shuffling tests correspondence.

structure over its retrained null. Routing, low-level and generation-control analyses retain their mixed outcomes, with native MindEye as a system reference. Our contribution is a measured-geometry readout and evidence linking its correct correspondence to image fidelity, while distinguishing those controlled effects from broad system superiority.

2 Related work

Generative reconstruction. Latent-diffusion decoding and Brain-Diffuser combine predictions at different visual levels (Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023). MindEye establishes contrastive alignment, a diffusion prior, and a low-level initialization path; MindEye2 extends shared-subject training and adaptation (Scotti et al., 2023; 2024). NeuroPictor combines pretraining with multilevel modulation, and SynMind supplies semantic constraints (Huo et al., 2024; Yang et al., 2026). We retain the image generator and selector within controlled comparisons and change the information available to neural readout.

Spatial readouts and measured correspondence. IRFA learns spatial and feature-selective assignments from macaque multi-unit responses (Le et al., 2025). Brain-IT uses functional voxel-cluster tokens to predict localized CLIP and VGG features (Beliy et al., 2026). These methods establish learned local readouts; pRF-Guide instead supplies measured pRF footprints and tests their correct correspondence at fixed capacity. Table 1 contrasts mechanisms, not performance under matched protocols.

Footprints and local prediction. Gaussian and compressive spatial-summation models characterize visual position and extent; Bayesian pRF models address uncertainty (Dumoulin & Wandell, 2008; Kay et al., 2013; Zeidman et al., 2018). We use the released effective size, and separately examine matching footprint extent to frozen VAE/CLIP sensitivities (Rombach et al., 2022; Radford et al., 2021). Calibration and selective prediction provide tools for testing local confidence (Guo et al., 2017; Geifman & El-Yaniv, 2017); our analysis distinguishes probability accuracy from the image changes caused by fusion.

3 pRF-Guide

Given averaged voxel responses $x \in \mathbb{R}^V$, pRF-Guide combines learned content with measured spatial preferences on a 16×16 visual grid. Figure 1 shows the readout and its semantic and low-level pathways. The primary controls change the prior while preserving the learned architecture.

3.1 Measured geometry in learned attention

For voxel v, Gaussian mass m_{vj} integrates its pRF center μ_v and released effective size s_v over cell j. Normalizing over cells within each voxel, then over voxels within each cell, gives K_{vj}. The $J = 256$ cells cover an 8.4° visual field in image-aligned coordinates. The spatial bias is

$$b_{vj} = 2\gamma_v \log R_{vj}, \qquad R_{vj} = \max\left\{ \frac{K_{vj}}{J^{-1} \sum_k K_{vk}}, 10^{-6} \right\}. \tag{1}$$

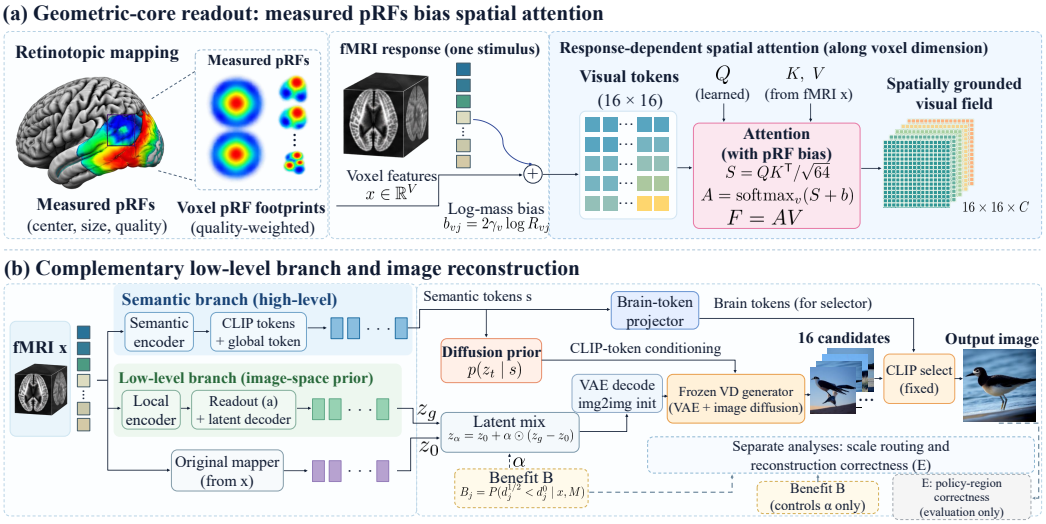


Figure 1: pRF-Guide overview. (a) Quality-weighted pRF footprints bias attention logits over response features. (b) Semantic tokens condition a diffusion prior, while independently predicted local and original-mapper latents are mixed and decoded for frozen image generation. Brain-projector features select among 16 candidates. The dashed benefit gate, scale routing and reconstruction correctness are separate analyses. Illustrations are schematic; Fig. 2 shows actual reconstructions.

Here $\gamma_v = \mathrm{clip}(q_v r_v, 0, 1)$. Quality q_v is the geometric mean of percentile ranks of pRF fit R^2 and response SNR; r_v weights the declared visual ROIs. The core study uses released SNR, and the expanded study estimates it from actual fitting-set repeats. Invalid fits receive a uniform footprint and zero geometric weight. Equation 1 expresses each voxel's relative spatial preference and attenuates it when measurement quality is low.

A shared scalar projection, voxel identity embedding and feature transform produce $h_v \in \mathbb{R}^{256}$. Learned queries and response-dependent keys give $c_{jv} = Q_j^\top W_K h_v / \sqrt{64}$. Attention then forms

$$A_{jv} = \mathrm{softmax}_v(c_{jv} + b_{vj}) = \frac{e^{c_{jv}} R_{vj}^{2\gamma_v}}{\sum_u e^{c_{ju}} R_{uj}^{2\gamma_u}}, \qquad F_j = \sum_v A_{jv} h_v. \tag{2}$$

Equation 2 reweights content scores without forcing a one-to-one assignment. Measured geometry adds no trainable parameters. *Learned attention* sets $b = 0$; the *shuffled pRF prior* permutes footprint rows within ROI, hemisphere and quality strata, retaining response features and quality at their original voxels, and is retrained. The true-minus-shuffled contrast tests correct correspondence under that permutation, while shuffled-minus-absent geometry measures the structured-prior effect in the same architecture.

3.2 Visual Prediction and Generation

The semantic readout predicts 256 CLIP patch tokens of dimension 768 and a global token (Radford et al., 2021). The core global token pools voxel features; the expanded model uses a separate global query. Its projector and MindEye diffusion prior are trained jointly with soft CLIP contrast, prior diffusion, patch cosine and patch contrastive losses weighted $1, 30, 5, 1$ (Scotti et al., 2023).

An independent local encoder predicts 64 channels on the grid, decoded to a VAE latent $z_g \in \mathbb{R}^{4\times64\times64}$. With the original mapper's prediction z_0, fusion is

$$z_\alpha = z_0 + \alpha \odot (z_g - z_0). \tag{3}$$

Zero fusion retains the original low-level path; half fusion sets $\alpha = 0.5$. The expanded branch trains through the half-blend using latent L_1, local contrastive and frozen-CLIP semantic losses, initializing its decoder from the native mapper fitted on the same neural data.

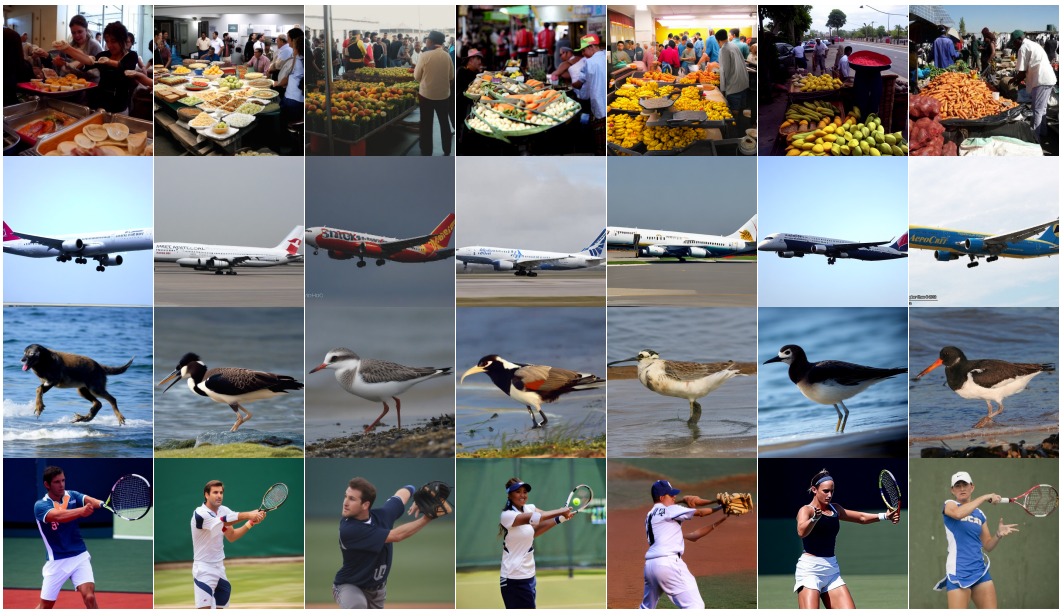


Figure 2: Selected Participant 1 test examples, in row order COCO73k 2950, 14610, 14898, and 43156. "Ours" is the historical geometric-core seed-42 checkpoint with the original low-level mapper. Other columns use released system outputs, including refined MindEye2 Participant 1 outputs; they retain different training and generation protocols.

Scale routing is a separate tested operator. For branch $\ell \in \{L, P, G\}$, set $w_{v\ell} = (1 - \gamma_v)/3 + \gamma_v \, \mathrm{softmax}_\ell[-(\log s_v^{\mathrm{eff}} - \log \rho_\ell)^2]$ and add $\log(w_{v\ell} + 10^{-6})$ to attention logits. The radii compare pRF and frozen-teacher sensitivity under a common cell-energy convention; the global query has no position bias. Uniform weights disable routing. Joint shuffling permutes footprint and routing size together, then refits both branches; size-only shuffling changes routing sizes alone. Fixed-weight center/width interventions retain the original routing sizes.

For the local-control analysis, $B_j = P(d_j^{1/2} < d_j^0 \mid x, \mathcal{M})$ predicts whether half fusion improves regional error $d_j = 1 - \mathrm{GridSSIM}_j$, with $\mathcal{M}$ denoting fixed metadata and fitted model parameters. The coefficient is $a_j = 0.5[2B_j - 1]_+$, bilinearly upsampled for Eq. 3. Separate calibration sets fit positive temperatures. A correctness head E reports the final policy's regional accuracy and affects neither fusion nor candidate selection. Full head and scale-estimation specifications are in Appendix A.4.

Frozen Versatile Diffusion receives predicted prior tokens and decoded low-level img2img initialization (Xu et al., 2023). We draw 16 candidates using 100 prior and 20 image steps, guidance 3.5 and strength 0.85. The original MindEye selector compares brain-projector and candidate CLIP features. Inference uses fMRI and metadata, with no target image. The core comparison shares its generator, low-level mapper and selector within each participant.

4 Experiments

4.1 Studies, controls, and outcomes

The geometric-core study uses NSD participants 1, 2, 5 and 7 (the original dataset identifiers), seeds 42, 123, and 2027, and three matched conditions, giving 36 runs (Allen et al., 2022; Prince et al., 2022). Each uses 8,559 training, 300 validation, and 982 test images. The expanded Participant 1, seed 42 study uses 7,959 neural-fitting, 300 validation, and 600 calibration images, with 27 evaluated output conditions on the same 982 test identities. Its retrained controls and fixed-weight interventions

Table 2: Controlled geometric-core comparison. Four NSD participants (1,2,5,7), three training seeds each, and 982 test images per run. Entries are participant means ± between-participant SD after averaging seeds. All rows share architecture, objectives and generation within a participant.

Readout	PixCorr↑	SSIM↑	CLIP↑	Spatial↑	Center↓	Brain↑
Learned attention	$.284 \pm .059$	$.313 \pm .007$	$.869 \pm .013$	$.566 \pm .007$	$.128 \pm .008$	$.345 \pm .049$
Shuffled pRF prior	$.289 \pm .062$	$.317 \pm .009$	$.859 \pm .027$	$.573 \pm .009$	$.112 \pm .010$	$.354 \pm .053$
pRF-Guide (core)	$.293 \pm .063$	$.319 \pm .008$	$.872 \pm .034$	$.577 \pm .012$	$.109 \pm .010$	$.356 \pm .053$

CLIP is two-way identification; Spatial is reconstructed-image SwAV spatial cosine; Center is energy-centroid error in degrees; Brain is auxiliary image-to-fMRI response correlation.

Table 3: External-system context on the same 982 test images from NSD Participant 1, evaluated by common code. Published systems retain their native training and sampling settings. Our core checkpoint (training seed 42) is the one illustrated in Fig. 2; the expanded refit in Table 4 uses a different fitting partition and architecture.

Method	PixCorr↑	SSIM↑	CLIP↑	Spatial↑	Center↓
MindEye (Scotti et al., 2023)	0.388	0.336	0.949	0.602	0.092
MindEye2 (Scotti et al., 2024)	0.325	0.358	0.943	0.613	0.086
Brain-IT (Beliy et al., 2026)	0.438	0.452	0.972	0.649	0.073
SynMind (Yang et al., 2026)	0.312	0.366	0.970	0.605	0.086
NeuroPictor (Huo et al., 2024)	0.272	0.354	0.944	0.622	0.079
pRF-Guide (core)	0.375	0.326	0.894	0.590	0.099

are distinguished below. Architectures, fitting partitions, and quality estimates differ from the core study. All comparisons use a benchmark accessed during earlier development.

Primary geometry outcomes are aligned SwAV spatial-feature cosine and feature-energy centroid error in degrees, computed from selected generated images and targets. PixCorr/SSIM measure appearance, CLIP two-way identification measures semantic discriminability, and GridSSIM averages RGB structural similarity over 256 nonoverlapping regions. Historical COCO-box SSIM uses grayscale object crops. An auxiliary image-to-fMRI ridge encoder measures brain-response consistency. These diagnostics characterize generated images rather than verify internal attention assignments (Appendix I).

Core summaries average three seeds within a participant and then weight four participants equally; SDs describe participant means. Expanded-study intervals use 10,000 paired image resamples with seed 20260923, averaging regions within images first. They are nominal 95% intervals, unadjusted for multiple comparisons and conditional on Participant 1, seed 42. The omission, regional and supervision-interaction summaries are post-hoc analyses of fixed outputs. The additional joint-null, routing, center and width controls are separately executed comparisons on this same test pool, not additional participants or training-seed replications.

4.2 The increment from correct correspondence

True geometry improves spatial cosine over the core retrained null by 0.00390 and reduces centroid error by 0.00272° (Figure 3). Both directions hold in all four participant means; favorable seed-pair counts are 10/12 and 11/12. Participant 7 has the weakest mean cosine increment, 0.00021, with two slightly negative seed contrasts. Leaving out each participant in turn retains cosine increments of 0.00287–0.00513 and centroid-error reductions of 0.00207–0.00316°. These are descriptive sensitivity summaries of the original runs (Appendix G).

The shuffled prior itself improves cosine over learned attention by 0.00696 and reduces centroid error by 0.01539°. Adding the correspondence increment gives total effects over learned attention of 0.01086 and -0.01810°, favorable in all twelve seed pairs. The larger shuffled-prior contrast shows that correct retinotopy does not explain the entire gain over geometry-free attention. The decomposition is arithmetic, not an allocation to independent biological causes.

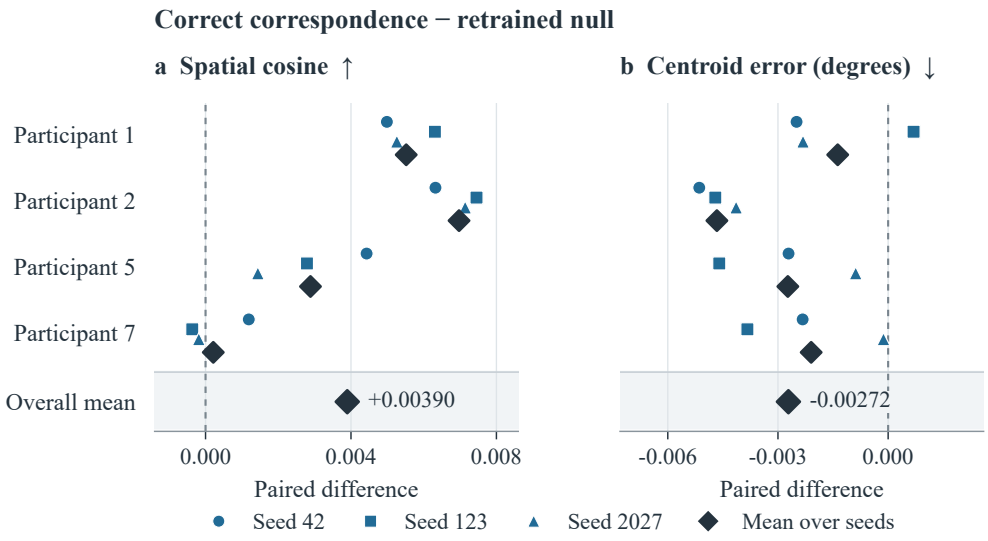


Figure 3: Correct correspondence versus the retrained shuffled prior. Marker shapes distinguish training seeds within each NSD participant; dark diamonds average three seeds. The shaded row gives the equally weighted participant mean. Vertical offsets separate markers, not uncertainty. Higher spatial cosine and lower centroid error favor correct correspondence.

Table 4: Expanded Participant 1 study, seed 42, 982 test images. Ours (expanded) is full geometry with half-fusion. The three prior controls also use half-fusion; the last row changes only the low-level readout. Native MindEye uses its original high-level architecture. Spatial, Center (degrees), and Brain follow Table 2.

Method	PixCorr ↑	GridSSIM ↑	CLIP ↑	Spatial ↑	Center ↓	Brain ↑
Native MindEye	0.3420	0.3357	0.9273	0.5884	0.1006	0.3408
No footprint or routing	0.2985	0.3250	0.8697	0.5623	0.1240	0.3198
Routing only	0.3010	0.3295	0.8598	0.5639	0.1285	0.3195
Footprint only	0.3416	0.3272	0.8787	0.5829	0.0999	0.3376
pRF-Guide (expanded)	0.3444	0.3263	0.8744	0.5825	0.0968	0.3381
Ours, original low ($\alpha = 0$)	0.3316	0.3195	0.8742	0.5802	0.1042	0.3329
Ours, geometry-free low	0.3161	0.3351	0.8781	0.5796	0.0973	0.3309

True geometry also improves feature-energy IoU and auxiliary brain-response correlation over the null, by 0.00595 and 0.00194. Relative to learned attention, mean COCO-box SSIM decreases by 0.00256 despite higher global SSIM, and Participant 7 CLIP identification decreases from 0.8516 to 0.8212. Candidate-mean images favor true geometry over both controls on all eight aggregated global measures, although their semantic identification is lower than that of selected images. Appendix J preserves these output-policy and participant outcomes.

4.3 Geometry correspondence and fixed-weight interventions

The expanded factorial comparison toggles footprint bias and target-scale routing in both neural branches, using fixed half-fusion. The footprint improves spatial cosine with routing by 0.01862 $[0.01687, 0.02040]$ and without routing by 0.02063 $[0.01881, 0.02244]$; centroid errors decrease in both settings. Native MindEye retains its own architecture and objective and serves as a system reference (Table 4).

To test measured correspondence beyond removing the prior, the joint-null control permutes each footprint and its routing reference size together within the declared strata, then retrains both

Table 5: Geometry controls in the expanded Participant 1, seed 42 study. Differences are full-half minus control on 982 images. Spatial cosine and GridSSIM favor positive differences; centroid error favors negative differences. Paired intervals are in Appendix B; full image metrics are in the standalone supplement.

Control	Type	Δ spatial	Δ centroid (°)	Δ GridSSIM
Center $x + 0.525°$	Inference	$+0.004017$	-0.003788	$+0.001073$
Footprint width $\times 2$	Inference	$+0.019205$	-0.011785	$+0.009427$
Routing-size shuffle	Retrained	$+0.001509$	$+0.000649$	$+0.003225$
Joint geometry shuffle	Retrained	$+0.009360$	-0.005922	$+0.005664$
Low-branch joint shuffle	Low swap	$+0.002322$	-0.002701	-0.000895

Inference controls retain learned weights and routing reference sizes. Retrained controls refit both branches. Low swap fixes the full high-level model and reuses the joint-shuffle low-level checkpoint. All use fixed half fusion.

neural branches. Relative to this matched control, full half-fusion improves spatial cosine by 0.00936 $[0.00775, 0.01099]$, reduces centroid error by $0.00592°$ $[-0.00999, -0.00179]$, and improves GridSSIM by 0.00566 $[0.00236, 0.00890]$. This extends the correspondence evidence to regional RGB structure. It concerns the joint footprint–routing assignment in this expanded model; the earlier core null is a separate protocol, not another shuffle replicate of this comparison.

Two fixed-weight interventions further test whether the fitted model uses specific geometric attributes. Translating all pRF centers by $+0.525°$ horizontally, with vertical coordinates unchanged, lowers spatial cosine by 0.00402 relative to the measured centers. Doubling footprint widths lowers cosine by 0.01920 and GridSSIM by 0.00943 $[0.00602, 0.01292]$ when differences are expressed as full minus perturbed. Both interventions retain the learned high/low weights, original routing reference sizes, native low-level mapper, and generation budget. They demonstrate sensitivity under changed geometry; distribution shift and the single tested dose limit any claim of globally optimal pRF parameters. Appendix B records the exact operators and all outcomes.

4.4 Routing and low-level contributions remain conditional

Retraining after shuffling only routing sizes gives smaller full-minus-shuffle increments: cosine 0.00151 $[0.00010, 0.00297]$ and GridSSIM 0.00322 $[0.00039, 0.00610]$. Yet adding routing to an existing footprint has no resolved cosine gain over routing-off, and the shuffled-size model has slightly higher PixCorr and CLIP identification than full. Correct allocation can outperform this random allocation without routing being necessary or uniformly beneficial. Both controls remain in Tables 4 and 5.

Holding full high-level conditioning fixed while substituting the same joint-null low-level checkpoint gives a full-minus-control cosine increment of 0.00232, but GridSSIM changes by -0.00089 $[-0.00334, 0.00158]$. This tests the low-level path without adding a new independently trained null. Against a geometry-free low-level branch, true geometry improves PixCorr and spatial cosine but reduces GridSSIM by 0.00885 $[-0.01171, -0.00605]$, with decreases in 61.6% of image means.

Static regional analyses locate part of this cost. For the routed footprint contrast, the highest coverage third has less GridSSIM loss than the lowest by 0.00560 $[0.00168, 0.00957]$, without a resolved positive high-coverage effect. Geometric low-level readout loses GridSSIM in all three coverage thirds and more strongly in outer regions. Coverage and eccentricity correlate at -0.577, so these are associated spatial descriptions, not isolated quality effects. All groups and gain/loss distributions appear in Appendix G.2.

4.5 Supervision protocols and generation control

Four executed high-level supervision budgets use nested subsets of 795, 1,989, 3,979 and 7,959 images (10%, 25%, 50%, 100%). All retain the full-data original low-level mapper at $\alpha = 0$; normalization, quality and teacher scales also use full T. Figure 4 compares full geometry with the matched geometry-free readout. Spatial cosine improves and centroid error decreases at every budget. At 25% and 50%, cosine differences are 0.01397 $[0.01214, 0.01580]$ and 0.01314 $[0.01142, 0.01489]$, while centroid changes are $-0.02741°$ $[-0.03306, -0.02186]$ and $-0.02439°$ $[-0.02956, -0.01921]$.

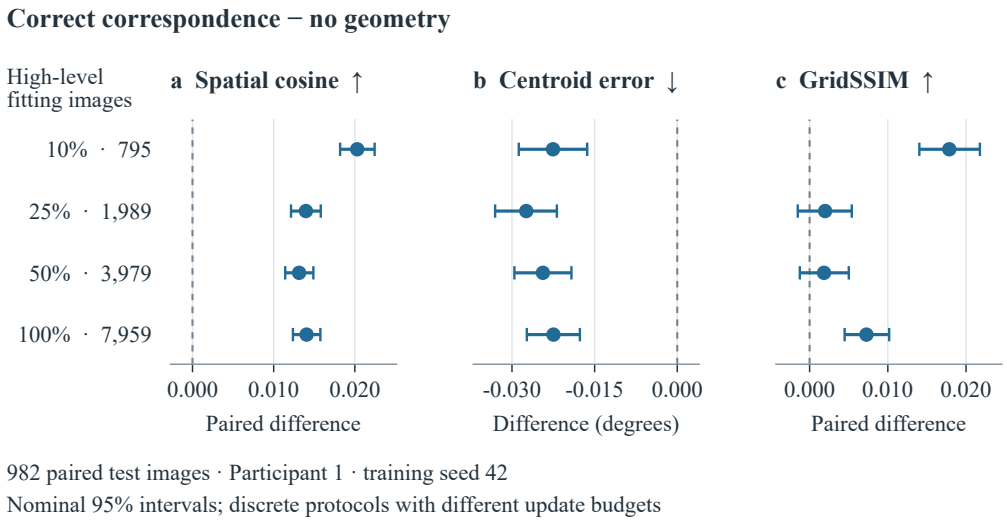


Figure 4: Measured geometry minus the matched geometry-free readout at four high-level supervision budgets. Points and bars are paired-image effects and nominal 95% intervals for NSD Participant 1, seed 42, using the same 982 test images. Lower centroid error is better. All budgets train for 120 epochs, with different update counts; this is not a fixed-optimization comparison. Native 50% is unavailable.

Local structure is less consistent: GridSSIM differences at 25% and 50% are 0.00199 and 0.00184, with intervals crossing zero. The earlier 10%-versus 100% GridSSIM interaction remains 0.01059 $[0.00600, 0.01514]$. All budgets train for 120 epochs, so update counts and selected checkpoints differ; this tests executed protocols, not label count at fixed optimization. Native MindEye retains stronger spatial/semantic results at 25% (cosine 0.5732 versus 0.5634; CLIP 0.8749 versus 0.7238); native 50% was not evaluated. Appendix G.4 reports all available native comparisons and training counts.

Fixed half-fusion improves GridSSIM over the same high-level model's original low-level pathway by 0.00678 $[0.00394, 0.00958]$. Calibrating its benefit head lowers Brier score from 0.2824 to 0.2490, close to the prevalence baseline's 0.2500, while contracting the mean fusion coefficient from 0.09008 to 0.01442. The calibrated policy loses GridSSIM relative to raw gating and half-fusion, and has no resolved advantage over mean-matched or rotated gates. Correctness calibration improves on prevalence but not on the combined simple-risk baseline. The new geometry controls all use fixed half-fusion and do not change these gate conclusions (Appendix D).

Table 3 compares our core model with released MindEye, MindEye2, Brain-IT, SynMind and NeuroPictor outputs on the same 982 Participant 1 identities (Scotti et al., 2023; 2024; Beliy et al., 2026; Yang et al., 2026; Huo et al., 2024). Brain-IT leads several global and spatial measures. Their native training and sampling protocols differ from the matched internal controls. Figure 2 retains the supplied examples from the historical Participant 1 geometric-core checkpoint, including remaining identity and pose errors.

5 Discussion and Conclusion

Measured correspondence adds information to an attention architecture already capable of spatial assignment. The four-participant core separates its increment from the larger effect of introducing a structured shuffled prior. In the expanded Participant 1 interface, retrained joint shuffling reduces spatial feature alignment and regional RGB structure. Fixed-weight center and width interventions provide complementary evidence that the fitted model uses these geometric attributes. Together these controls strengthen the measured-correspondence interpretation without treating inference-time sensitivity as optimality or independent replication.

The utility of that information depends on how it enters generation. Correct routing sizes have a small advantage over shuffled sizes but no resolved gain over routing-off on the principal spatial outcome. Low-level geometry has metric-dependent costs, and better calibrated benefit probabilities do not improve the tested gate. The supervision comparisons span four executed budgets, with native MindEye retaining stronger semantic identification where evaluated. These results locate both useful correspondence and limits of the expanded interface.

The core study uses one shuffle per participant, and the expanded controls share Participant 1, seed 42 and one joint permutation. Image-bootstrap intervals do not measure variability across such repetitions. Spatial and auxiliary brain scores share SwAV features, while the joint-null GridSSIM contrast supplies a complementary RGB structural outcome. Within this scope, pRF-Guide demonstrates a contribution of measured retinotopy to learned reconstruction, supported by correspondence controls and geometric interventions while retaining its structural and semantic tradeoffs.

ETHICS STATEMENT

The experiments use the public Natural Scenes Dataset, acquired with participant consent and distributed in de-identified form (Allen et al., 2022). The task reconstructs images viewed during the recorded experiment. Participant identities are not inferred or disclosed. The supplied reconstructions illustrate model outputs and should not be interpreted as direct recordings of subjective visual experience.

REPRODUCIBILITY STATEMENT

The appendix specifies data partitions, measured geometry, optimization, generation, and metric definitions for both studies. The supplementary package includes the core model and evaluation code, configuration and checkpoint specifications, and the numerical result exports used to build the tables. Training and generation require the separately distributed NSD resource and pretrained visual models.

AI USE STATEMENT

Generative AI tools were used to aid and polish writing, including language editing and manuscript organization; for retrieval and discovery, including finding related work and locating relevant sources; and for research ideation and execution, including method specification, experiment planning and feedback, code inspection and analysis scripting.

REFERENCES

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

Roman Beliy, Amit Zalcher, Jonathan Kogman, Navve Wasserman, and Michal Irani. Brain-IT: Image reconstruction from fMRI via brain-interaction transformer. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c9fd326fd03eaf52f672c31cde9658af-Abstract-Conference.html.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html.

Serge O. Dumoulin and Brian A. Wandell. Population receptive field estimates in human visual cortex. *NeuroImage*, 39(2):647–660, 2008. doi: 10.1016/j.neuroimage.2007.09.034. URL https://doi.org/10.1016/j.neuroimage.2007.09.034.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Jingyang Huo, Yikai Wang, Yun Wang, Xuelin Qian, Chong Li, Yanwei Fu, and Jianfeng Feng. NeuroPictor: Refining fMRI-to-image reconstruction via multi-individual pretraining and multi-level modulation. In *Computer Vision – ECCV 2024*, pp. 56–73, 2024. doi: 10.1007/978-3-031-72983-6_4. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/06746.pdf.

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://doi.org/10.1152/jn.00105.2013.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012. URL https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Lynn Le, Thirza Dado, Katja Seeliger, Paolo Papale, Antonio Lozano, Pieter Roelfsema, Yağmur Güçlütürk, Marcel van Gerven, and Umut Güçlü. Inverse receptive field attention for naturalistic image reconstruction from the brain. *arXiv preprint arXiv:2501.03051*, 2025. doi: 10.48550/arXiv.2501.03051. URL https://arxiv.org/abs/2501.03051.

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, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.html.

Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13:15666, 2023. doi: 10.1038/s41598-023-42891-8. URL https://www.nature.com/articles/s41598-023-42891-8.

Jacob S. Prince, Ian Charest, Jan W. Kurzawski, John A. Pyles, Michael J. Tarr, and Kendrick N. Kay. Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *eLife*, 11:e77599, 2022. doi: 10.7554/eLife.77599. URL https://elifesciences.org/articles/77599.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Paul S. Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J. Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, and Tanishq Mathew Abraham. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. In *Advances in Neural Information Processing Systems*, volume 36, pp. 24705–24728, 2023. doi: 10.52202/075280-1073. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4ddab70bf41ffe5d423840644d3357f4-Abstract-Conference.html.

Paul Steven Scotti, Mihir Tripathy, Cesar Torrico, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. MindEye2: Shared-subject models enable fMRI-to-image with 1 hour of data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44038–44059, 2024. URL https://proceedings.mlr.press/v235/scotti24a.html.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14453–14463, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Takagi_High-Resolution_Image_Reconstruction_With_Latent_Diffusion_Models_From_Human_Brain_CVPR_2023_paper.html.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114, 2019. URL https://proceedings.mlr.press/v97/tan19a.html.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861. URL https://doi.org/10.1109/TIP.2003.819861.

Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7754–7765, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Xu_Versatile_Diffusion_Text_Images_and_Variations_All_in_One_Diffusion_ICCV_2023_paper.html.

Lan Yang, Minghan Yang, Ke Li, Honggang Zhang, Kaiyue Pang, and Yi-Zhe Song. SynMind: Reducing semantic hallucination in fMRI-based image reconstruction. *arXiv preprint arXiv:2601.17857*, 2026. doi: 10.48550/arXiv.2601.17857. URL https://arxiv.org/abs/2601.17857.

Peter Zeidman, Edward Harry Silson, Dietrich Samuel Schwarzkopf, Chris Ian Baker, and Will Penny. Bayesian population receptive field modelling. *NeuroImage*, 180:173–187, 2018. doi: 10.1016/j.neuroimage.2017.09.008. URL https://doi.org/10.1016/j.neuroimage.2017.09.008.

A EXPANDED-STUDY IMPLEMENTATION

The expanded interface uses NSD Participant 1 and training seed 42, with independent local and semantic encoders. Its global attention query, routing term and training-derived reliability distinguish it from the geometric core. The standalone supplementary PDF gives the complete executed specification; this appendix records the settings needed to interpret the comparisons.

A.1 PARTITIONS AND MEASUREMENT QUALITY

Split seed 20260924 divides the 8,859 non-test identities into 7,959 fitting images T, 300 validation images V, and 600 calibration images. All repeated responses to one image remain together. The calibration partitions contain 200 images for endpoint-risk/benefit fitting, 100 for benefit temperature, 150 for final-correctness fitting and 150 for correctness temperature. The seven partitions, including 982 test identities, are pairwise disjoint. Feature normalization uses T. Prior benchmark access and the exploratory status of the added simple-risk baselines are preserved.

The reliability estimate pools within-image variance over actual response repeats, subtracts the repeat-count-adjusted noise contribution from between-image variance, clips signal variance at zero, and takes the square root of the signal-to-noise variance ratio. Average-tie percentile ranks of this SNR and pRF fit R^2 define q_v by their geometric mean. The fitting set contains 6,799 images with three repeats, 873 with two and 287 with one; averaged padding is excluded. Of 15,724 voxels, 12,799 have valid joint quality/geometry estimates. Invalid voxels retain responses but receive zero geometric reliability. ROI weights are 1 for labels 1–7 and 0.5 otherwise. Coordinates are $x = e\cos\theta$, $y = -e\sin\theta$, over an 8.4° field.

A.2 SPATIAL SCALES AND MODEL FITTING

Teacher radii use 64 fitting images, seed 20260923, a 32×32 perturbation grid, two fixed Rademacher RGB directions per cell, and paired ± 0.02 perturbations clipped to $[0, 1]$. Normalize squared finite-difference responses over input cells and compute the square root of half their centered spatial second moment. pRF radii apply the same moment to normalized squared Gaussian cell masses before cross-voxel normalization. The branch medians are $\rho_L = 0.21567^\circ$, $\rho_P = 1.86779^\circ$ and $\rho_G = 2.44312^\circ$, with no forced ordering. VAE sensitivity measurement uses bicubic 512-pixel resizing, whereas latent-target fitting uses bilinear resizing; the radii are effective sensitivity estimates under the measurement preprocessing.

The encoders use 256-dimensional voxel features and 64-dimensional attention keys/queries. The semantic projector is $768 \to 2048 \to 2048 \to 768$; the diffusion prior has six layers, 12 heads, head dimension 64, 100 steps and conditioning dropout 0.2. High-level training uses 120 epochs, microbatch 8, effective batch 32, AdamW peak learning rate 3×10^{-4} and weight decay 10^{-2}. Validation uses fixed diffusion noise for model selection. Native MindEye keeps its original mapper and objective.

The native low-level mapper trains for 120 epochs. The geometric branch initializes its decoder from that mapper and trains for 20 epochs with microbatch 2, effective batch 32, AdamW learning rate 10^{-4}, weight decay 10^{-2} and a cosine schedule. Its half-blend objective combines latent $L_1/0.18215$, ConvNeXt local contrastive loss weighted 0.1, and real-image CLIP token cosine loss weighted 0.5. Gradients through frozen VAE/CLIP teachers update the local branch.

A.3 PROBABILITY HEADS AND AUXILIARY EVALUATION

The regional base input has 590 dimensions: local projection 64, semantic patch 256, broadcast global 256, coverage/quality 2, and latent summaries 12. The risk head has hidden widths 64/32 and two softplus outputs. Benefit appends its two predictions (592 inputs); correctness instead appends gate 1 and fused latent 4 (595 inputs), without endpoint risks. Binary heads have hidden width 32. Heads use 200 epochs, batch 8, AdamW learning rate 10^{-3} and weight decay 10^{-4}, with equal image weights and normalization fitted on the benefit-fitting partition. Fitted temperatures are $T_B = 7.33864$ and $T_E = 2.15076$; the validation error threshold is $\eta = 0.77963036$. Correctness never alters the gate or candidate selection.

GridSSIM uses RGB 512 images, 256 nonoverlapping 32×32 regions, an 11×11 Gaussian window, $\sigma = 1.5$, population moments and data range 1. Targets are bicubic-antialiased and clamped before scoring saved selected reconstructions (Wang et al., 2004). The independent ridge evaluator uses SwAV average-pool features fitted on 8,259 historical training identities, with 300 for penalty selection; it is excluded from reconstruction fitting and selection. This auxiliary evaluator does not use the expanded decoder's T partition. SwAV, AlexNet, Inception, EfficientNet and CLIP define the image metrics; COCO supplies object annotations (Caron et al., 2020; Krizhevsky et al., 2012; Szegedy et al., 2016; Tan & Le, 2019; Radford et al., 2021; Lin et al., 2014; Liu et al., 2022).

A.4 Routing and local-control definitions

For target branch $\ell \in \{L, P, G\}$, the scale allocation is

$$w_{v\ell} = \frac{1 - \gamma_v}{3} + \gamma_v \frac{\exp[-(\log s_v^{\text{eff}} - \log \rho_\ell)^2]}{\sum_k \exp[-(\log s_v^{\text{eff}} - \log \rho_k)^2]}. \tag{4}$$

Each readout adds $\log(w_{v\ell} + 10^{-6})$ to its attention logits; the global query has no spatial bias. Routing-off sets $w_{v\ell} = 1/3$. Footprint-off sets the spatial kernel to uniform. The two interventions are applied in both neural branches. Footprint width remains part of the spatial operator, distinct from cross-target allocation.

Let $d_j = 1 - \text{GridSSIM}_j^{\text{RGB32}}$ be error of the selected reconstruction. With fixed metadata and model parameters $\mathcal{M}$, the local probability targets are

$$B_j(x) = P(d_j^{1/2} < d_j^0 \mid x, \mathcal{M}), \qquad a_j = \tfrac{1}{2}[2B_j - 1]_+, \tag{5}$$

$$E_j(x) = P(d_j(\hat{I}_a, I) \le \eta \mid x, \mathcal{M}, a). \tag{6}$$

The preceding protocol specifies separate fitting and temperature sets, exact input features, bilinear upsampling and the correctness threshold. E reports the frozen policy's outcome; it does not set the fusion coefficient or select a candidate. Positive-temperature calibration changes the strength assigned to a positive benefit logit but preserves its order. The standalone supplement gives the expanded definitions and coefficient analysis.

B Joint correspondence and geometric interventions

The five additional controls share Participant 1, training seed 42, the T/V/calibration split, native low-level mapper, and 982 test identities with the expanded full model. Every comparison in this section uses fixed $\alpha = 0.5$, 16 candidates, 100 prior steps, 20 image-diffusion steps, guidance 3.5, strength 0.85, and the original candidate selector. They evaluate geometry at fixed half-fusion and do not test calibrated benefit gating. The stored arrays preserve identical ordered image IDs across the reference and all controls.

Retrained correspondence controls. Permutation seed 2027 generates a bijection within ROI, hemisphere, and positive-versus-zero quality strata. Of 15,724 voxels, 15,688 change index. Quality and ROI weights remain attached to their original responses. The joint null permutes the footprint and effective routing reference size as a pair. The routing-size shuffle changes the latter alone, preserving each measured footprint. Both controls refit high-level and geometric low-level branches with the same architectures, objectives and budgets as full: 120 high-level epochs and 20 low-level epochs. The low-level decoders use the same native-mapper initialization. Validation selects checkpoints under the same rules.

The low-only joint-null condition combines the full high-level checkpoint with the very same joint-null low-level checkpoint used in the joint comparison. It is a branch substitution, not another independent null training. The observed high/low combinations do not identify a full interaction surface or justify assigning a percentage of the overall gain to either branch. The historical four-participant null uses a different training split and core readout and is not pooled with this expanded-model permutation as an additional null replicate.

Table 6: Paired full-half-minus-control differences for all five completed geometry controls, with nominal 95% image-bootstrap intervals. Resampling uses 982 paired images, 10,000 draws, and seed 20260923. Intervals are conditional on Participant 1, seed 42 and are unadjusted across comparisons.

Control	Δ spatial	Δ centroid (°)	Δ GridSSIM
Center $x + 0.525°$	$+0.004017$ $[+0.002746, +0.005291]$	-0.003788 $[-0.007335, -0.000179]$	$+0.001073$ $[-0.001540, +0.003650]$
Footprint width $\times 2$	$+0.019205$ $[+0.017426, +0.021058]$	-0.011785 $[-0.015998, -0.007644]$	$+0.009427$ $[+0.006022, +0.012923]$
Routing-size shuffle	$+0.001509$ $[+0.000096, +0.002968]$	$+0.000649$ $[-0.003055, +0.004404]$	$+0.003225$ $[+0.000394, +0.006097]$
Joint geometry shuffle	$+0.009360$ $[+0.007751, +0.010989]$	-0.005922 $[-0.009994, -0.001786]$	$+0.005664$ $[+0.002359, +0.008899]$
Low-branch joint shuffle	$+0.002322$ $[+0.001155, +0.003485]$	-0.002701 $[-0.006170, +0.000708]$	-0.000895 $[-0.003341, +0.001580]$

Fixed-weight geometric interventions. The center intervention adds $0.525°$ to every horizontal center coordinate, leaving the vertical coordinate unchanged. This equals one cell of the 16×16 readout grid. The width intervention multiplies the raw Gaussian footprint width by two. Each replaces the position-dependent buffers in both fitted branches while retaining their learned parameters, native low-level mapper, quality/ROI weights and teacher-scale estimates. The effective routing reference sizes and routing weights are held fixed. Thus width perturbation acts on spatial footprint extent rather than cross-target allocation. These are single-dose inference interventions; sensitivity can include a mismatch with the training geometry and does not imply that the measured parameters are globally optimal.

Evaluation and scope. For each control, we subtract its per-image metric from the full-half reference, averaging the 256 GridSSIM regions within images first. The same 982 image indices are resampled in pairs for 10,000 draws with seed 20260923. Percentile intervals are nominal 95%, unadjusted across the fifteen primary contrasts. Table 6 retains all intervals, including those crossing zero; the standalone supplement provides the full metric surface. The source results and controls are retrospective evaluations on the existing test pool. Completed 25% and 50% supervision conditions are reported separately in Appendix G.4.

C Factorial geometry analysis

The factorial conditions are full (footprint and routing), footprint only, routing only, and neither. Both branches receive the designated geometry buffers and are fitted under the same expanded protocol. The low-level coefficient is fixed at 0.5. Each conditional effect compares matched architectures; the interaction is the routing effect with the footprint minus the routing effect without it. Table 7 reports the spatial and structural effects on selected generated images. These downstream measures do not identify which internal attention weights are correct.

D Fusion policies and probability evaluation

All policies below share the full high-level model, geometric low-level branch, native low-level mapper, and candidate budget. The correctness head reports the calibrated policy's errors and does not choose candidates. Table 8 preserves the tested gates; their complete paired structural intervals are given in the standalone supplement. Calibration contracts the mean coefficient and spatial variance, while mean matching and rotation isolate gate magnitude from alignment. These diagnostics describe the resulting policy rather than establish a unique causal explanation for its performance.

D.1 Benefit and correctness probabilities

Brier scores and expected calibration error (ECE, 15 equal-width probability bins) use all 256 regions with equal image weight. The prevalence predictor is estimated on the appropriate head-

Table 7: Conditional footprint and routing effects. Effects and nominal 95% intervals refer to selected reconstructions, Participant 1, seed 42.

Added component	Metric	Δ	95% interval
Footprint, routing on	Spatial cosine	0.01862	[0.01687, 0.02040]
Footprint, routing on	Center error (°)	-0.03166	[-0.03678, -0.02656]
Footprint, routing on	GridSSIM	-0.00324	[-0.00674, 0.00019]
Footprint, routing off	Spatial cosine	0.02063	[0.01881, 0.02244]
Footprint, routing off	Center error (°)	-0.02414	[-0.02935, -0.01897]
Footprint, routing off	GridSSIM	0.00215	[-0.00145, 0.00569]
Routing, footprint on	Spatial cosine	-0.00041	[-0.00175, 0.00097]
Routing, footprint on	Center error (°)	-0.00303	[-0.00674, 0.00074]
Routing, footprint on	GridSSIM	-0.00088	[-0.00356, 0.00174]
Routing, footprint off	Spatial cosine	0.00160	[0.00009, 0.00307]
Routing, footprint off	Center error (°)	0.00449	[-0.00040, 0.00944]
Routing, footprint off	GridSSIM	0.00451	[0.00150, 0.00753]
Interaction	Spatial cosine	-0.00201	[-0.00402, 0.00000]
Interaction	Center error (°)	-0.00752	[-0.01383, -0.00120]
Interaction	GridSSIM	-0.00539	[-0.00942, -0.00139]

Table 8: Fusion policies with the same full high-level model and geometric low-level branch. All values are selected-image test scores; the correctness head never changes selection.

Method	PixCorr ↑	GridSSIM ↑	CLIP ↑	Spatial ↑	Center ↓	Brain ↑	Mean a
Full, original low ($\alpha = 0$)	0.3316	0.3195	0.8742	0.5802	0.1042	0.3329	0.0000
Full, half-fusion	0.3444	0.3263	0.8744	0.5825	0.0968	0.3381	0.5000
Coverage gate	0.3439	0.3251	0.8728	0.5826	0.0988	0.3353	0.4032
Risk gate	0.3374	0.3253	0.8722	0.5808	0.1013	0.3323	0.0874
Raw gate	0.3400	0.3254	0.8773	0.5818	0.0998	0.3346	0.0901
Calibrated gate	0.3334	0.3208	0.8745	0.5801	0.1037	0.3327	0.0144
Mean-matched gate	0.3337	0.3204	0.8765	0.5804	0.1036	0.3326	0.0144
Rotated gate	0.3341	0.3205	0.8772	0.5805	0.1037	0.3337	0.0144

fitting partition and evaluated on test. A positive temperature preserves the ordering of a head's logits. Raw-versus-calibrated comparisons therefore assess probability accuracy, while ranking comparisons require different predictors. Benefit calibration reduces Brier score by 0.38% relative to its fitting-prevalence baseline; correctness calibration reduces it by 23.96%.

Table 9: Held-out local probability scores. B predicts half-fusion improvement; E predicts correctness of the final calibrated policy. Lower Brier and ECE are better.

Head	Predictor	Brier	ECE	Mean probability	Event rate
B	Raw	0.28237	0.15281	0.50181	0.50918
B	Calibrated	0.24900	0.00853	0.50069	0.50918
B	Fit prevalence	0.24996	0.00628	0.51547	0.50918
E	Raw	0.20216	0.09822	0.48541	0.51406
E	Calibrated	0.19035	0.02088	0.50777	0.51406
E	Fit prevalence	0.25034	0.02315	0.49091	0.51406

The calibrated correctness head has Brier score 0.19035. Calibrated simple-risk baselines obtain 0.19055 from U_0, 0.19299 from $U_{1/2}$, and 0.18918 from their combination. Correctness-head minus combined-risk Brier is 0.00116 with interval $[-0.00141, 0.00380]$. Against U_0 it is -0.00020 $[-0.00286, 0.00247]$, and against $U_{1/2}$ it is -0.00265 $[-0.00528, 0.00001]$. None of these intervals resolves an advantage for the correctness head. The additional baselines are exploratory, use frozen model predictions, and leave both the original head and generated images unchanged.

E HIGH-LEVEL SUPERVISION ENDPOINTS

The endpoint comparison uses 795 (10%) and 7,959 (100%) high-level supervised identities. All methods retain the full-T native low-level mapper and set fusion to zero. Normalization, reliability, and target scales also use full T. The 100% anchors are the native refit, full zero-fusion model, and geometry-free high-level model with the same original mapper. Table 11 gives the endpoint and paired interaction results.

At 10%, full geometry improves spatial cosine, centroid error, and GridSSIM relative to matched geometry-free attention. Relative to native MindEye, it improves GridSSIM but lowers spatial cosine, CLIP identification, and brain-response correlation. The 25% and 50% extensions are reported separately; the native 50% model is unavailable. A 10% model can also exceed its 100% counterpart on one appearance metric while losing semantic discriminability; the full metric surface keeps these outcomes visible.

F Complete expanded-study metrics

The standalone supplement and numerical metadata report all global, spatial, local and auxiliary neural metrics for 27 evaluated conditions. The standalone supplement retains the full metric tables for the original 17 conditions; Appendix B gives the five geometry controls, and Appendix G.4 adds the five completed supervision conditions. "Neither prior" with half-fusion is distinct from the 100% geometry-free high-level model with the original mapper. The historical external-system table uses the earlier Participant 1 geometric-core checkpoint, not one of these expanded models.

G Reanalysis of frozen model outputs

These analyses reuse the same models, image identities and generated outputs as the reported core and expanded studies. Their purpose is to resolve the correspondence increment, spatial distribution of structural changes, and interaction between two supervision protocols. Comparisons and bins were fixed before computing these summaries, after earlier aggregate benchmark results had already been observed. They are post-hoc analyses, not new training replications or a prospective untouched test. All specified contrasts and bins are reported.

G.1 Correspondence increment and participant sensitivity

The two primary core measures are spatial feature cosine and centroid error. For every participant and training seed, we compute true minus null, null minus learned attention, and true minus learned attention. We then average the three seeds within a participant and weight participants equally. Leaving out each participant in turn gives four descriptive sensitivity summaries; these ranges are not confidence intervals. Table 10 reports participant summaries; all individual seed contrasts are retained in the standalone supplement. The sum of the first two contrasts equals the third, which is an arithmetic identity, not causal mediation. Four participants remain the independent units; no image-level uncertainty is reconstructed from saved run aggregates.

G.2 Regional structure, coverage, and eccentricity

The three contrasts are half minus footprint-off, routing-off minus both-off, and half minus low-only-both-off. The first two add footprints with and without routing in both neural branches. The third changes the low-level geometry while holding the full high-level model fixed. Regional GridSSIM arrays have shape 982×256 with the same ordered unique image IDs.

Static coverage is $C_j = \sum_v \gamma_v K_{vj}$ from the executed full-geometry operator using training-derived quality. Sort the 256 cells by coverage, breaking ties by row-major index, and partition into thirds containing 86, 85 and 85 cells. A separate radial partition uses cell-center eccentricity in $[0, 1.4)$, $[1.4, 2.8)$ and $[2.8, \infty)$ degrees, containing 24, 64 and 168 cells. Neither partition uses test errors to define its boundaries. Coverage and eccentricity have Pearson correlation -0.577 across grid cells; the separate partitions do not adjust for this association. Static coverage does not vary by image and is not an estimate of dynamic attention quality.

For each contrast, regions are averaged within each image and bin before equal image averaging. The 10,000 bootstrap samples use paired whole images with seed 20260923, preserving all regions within each sampled image. High-minus-low coverage and outer-minus-inner radial differences use the same image draws. Intervals are nominal and unadjusted across comparisons. Figure 5 reports all groups. The standalone supplement tabulates every interval and the image-level gain/loss frequencies, signed conditional means and quantiles.

Table 10: Participant-level correspondence contrasts in the historical geometric-core study. T is true pRF, N is the retrained within-stratum null, and L is matched learned attention. Positive cosine differences and negative centroid-error differences favor the first condition.

	Spatial cosine difference			Centroid-error difference (°)		
Participant / seed	$T-N$	$N-L$	$T-L$	$T-N$	$N-L$	$T-L$
Participant 1 seed mean	+0.00552	+0.00970	+0.01522	-0.00137	-0.01649	-0.01787
Participant 2 seed mean	+0.00697	+0.00729	+0.01426	-0.00466	-0.01886	-0.02352
Participant 5 seed mean	+0.00289	+0.00705	+0.00993	-0.00273	-0.01245	-0.01518
Participant 7 seed mean	+0.00021	+0.00383	+0.00404	-0.00210	-0.01374	-0.01584
Four-participant mean	+0.00390	+0.00696	+0.01086	-0.00272	-0.01539	-0.01810
Favorable seed pairs	10/12	12/12	12/12	11/12	12/12	12/12
Favorable participants	4/4	4/4	4/4	4/4	4/4	4/4
Mean excluding Participant 1	+0.00336	+0.00605	+0.00941	-0.00316	-0.01502	-0.01818
Mean excluding Participant 2	+0.00287	+0.00686	+0.00973	-0.00207	-0.01423	-0.01629
Mean excluding Participant 5	+0.00423	+0.00694	+0.01117	-0.00271	-0.01636	-0.01907
Mean excluding Participant 7	+0.00513	+0.00801	+0.01314	-0.00292	-0.01593	-0.01886

Participant means average seeds 42, 123, and 2027; the group mean weights participants equally. The last four rows average the remaining three participants. All contrasts use the same 36 fitted models and 982 selected test images per model. $(T-L) = (T-N) + (N-L)$ is an arithmetic identity.

G.3 Paired interaction of supervision protocols

For each image, average its 256 GridSSIM values in each of the four full/no-geometry, 10%/100% outputs, then form the difference of differences. Resample the same image indices for every output in a bootstrap draw. Native MindEye's two endpoints are carried through the same draws as contextual comparisons. Table 11 reports endpoint means, paired gaps and the interaction.

$$\Delta_{\text{int}} = (\overline{m}_{\text{full},10} - \overline{m}_{\text{no geom},10}) - (\overline{m}_{\text{full},100} - \overline{m}_{\text{no geom},100}) = 0.01059\ [0.00600, 0.01514]. \quad (7)$$

All six models run for 120 epochs. The 10% protocols have 25 optimizer-step opportunities per epoch, and the 100% protocols 249. Mixed-precision skipped steps give 2,996 realized updates in each 10% run, 29,865 in full/no-geometry 100% runs and 29,867 in native 100%. The one-cycle schedule scales with the total opportunities. Validation also selects different checkpoints: full/no-geometry 10% use zero-based epoch 111; full 100% uses epoch 118 and no-geometry 100% epoch 113. Native 10% and 100% select epochs 46 and 55. This is an interaction between executed supervision and optimization protocols, not a fixed-update estimate of sample-size causality.

G.4 Four completed high-level supervision budgets

The 25% extension adds matched geometry, geometry-free and native models with 1,989 fitting images; the 50% extension adds the two matched models with 3,979 images. Every available setting completes 120 training epochs, 982 test reconstructions and the full metric pipeline. The two matched models have identical fitting identities at each budget. All sets are nested 10%⊂25%⊂50%⊂100%, with no test overlap. The original 22 conditions' 594 shared scalar values are unchanged in the completed 27-condition snapshot.

The analysis pairs each geometry model with its same-budget reference and resamples whole image identities 10,000 times using seed 20260923. Regional values are averaged within image first. Native comparisons exist at 10%, 25% and 100%; no native 50% value is imputed. Table 12 records available means and optimization budgets; actual successful updates and selected-checkpoint counts are in the released supervision metadata. Table 13 preserves all principal intervals, including adverse native comparisons and the unresolved 25%/50% GridSSIM directions. The previously reported two-endpoint interaction is unchanged and remains a protocol interaction, not a fixed-update sample-efficiency result.

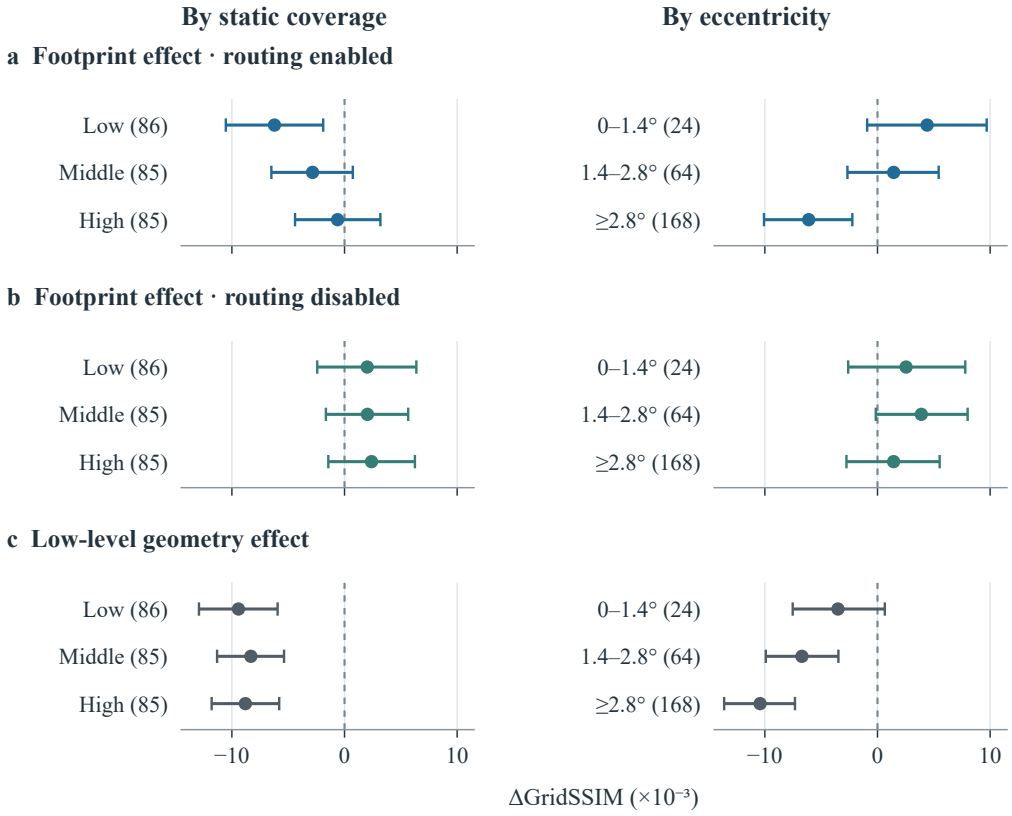


Figure 5: Regional GridSSIM changes, stratified separately by static coverage and eccentricity. Markers and bars show paired-image mean effects and nominal 95% intervals. Parentheses give the number of image-grid cells in each group. The routed footprint has a smaller structural loss at high coverage, without a resolved positive high-coverage effect; these associations do not isolate a causal quality contribution.

H HISTORICAL GEOMETRIC-CORE PROTOCOL

The core study uses NSD participants 1, 2, 5 and 7 with 8,559 training images, 300 validation images, and 982 test images per participant. Repeated responses are averaged within stimulus. Input widths are 15,724, 14,278, 13,039, and 12,682 voxels, respectively. Response arrays, pRF centers and widths, ROI labels, and quality weights retain a fixed voxel order. Three conditions and training seeds 42, 123, and 2027 give 36 trained models. Validation selects a checkpoint from each 120-epoch run. This study has no cross-target routing or probability head.

H.1 FOOTPRINT, QUALITY, AND CORRESPONDENCE CONTROLS

The 16×16 grid covers an $8.4°$ field, with cell width $0.525°$. For pRF center μ_v and released effective size s_v, exact rectangular-cell Gaussian mass is

$$m_{vj} = \prod_{a \in \{x,y\}} \left[\Phi\left(\frac{u_j^a - \mu_v^a}{s_v} \right) - \Phi\left(\frac{l_j^a - \mu_v^a}{s_v} \right) \right], \tag{8}$$

where l_j^a, u_j^a are the cell boundaries and Φ is the standard normal CDF. Row normalization within a voxel followed by column normalization across voxels gives K. Coordinates follow $x = e\cos\theta$, $y = -e\sin\theta$. Valid pRFs have finite centers, positive finite size and fit R^2, finite noise-ceiling SNR, eccentricity and size at most $8.4°$, and nonzero display mass. Invalid fits use a uniform footprint and zero geometric reliability.

Table 11: High-level supervision protocols and their paired GridSSIM interaction. All outcomes use Participant 1, seed 42, the same 982 test identities, the full-data original low-level mapper, and zero geometric fusion. Intervals are nominal 95% paired-image bootstrap intervals (10,000 draws).

High-level model	Images	GridSSIM [95% interval]	CLIP	Brain
Full geometry	795	0.32722 [0.31805, 0.33646]	0.6182	0.1749
No geometry	795	0.30936 [0.30110, 0.31758]	0.5687	0.1435
Full geometry	7,959	0.31951 [0.31009, 0.32881]	0.8742	0.3329
No geometry	7,959	0.31224 [0.30289, 0.32146]	0.8709	0.3235
Native MindEye	795	0.30760 [0.29880, 0.31623]	0.8327	0.3243
Native MindEye	7,959	0.33567 [0.32613, 0.34519]	0.9273	0.3408

Paired contrast	ΔGridSSIM	95% interval
Full minus no geometry (10%)	+0.01786	[+0.01403, +0.02177]
Full minus no geometry (100%)	+0.00727	[+0.00448, +0.01017]
Geometry gap at 10% minus gap at 100%	+0.01059	[+0.00600, +0.01514]
Full minus native (10%)	+0.01962	[+0.01627, +0.02301]
Full minus native (100%)	-0.01616	[-0.01871, -0.01354]
No geometry minus native (10%)	+0.00176	[-0.00192, +0.00541]
No geometry minus native (100%)	-0.02342	[-0.02630, -0.02051]

The 10% and 100% protocols both train for 120 epochs but allocate 3,000 and 29,880 optimizer-step opportunities, respectively. Their learning-rate schedules and selected checkpoints consequently differ. This is a two-protocol interaction, not an isolated label-count effect or a whole-system learning curve. CLIP identification and the auxiliary brain-response correlation are descriptive context; native MindEye uses its original architecture and objective.

Table 12: Executed high-level supervision budgets for Participant 1, training seed 42. All rows share the full-T native low-level mapper and $\alpha = 0$. Each evaluated high-level model completed 120 epochs; native at 50% was not evaluated. Center error is in degrees.

Model	Budget	PixCorr ↑	GridSSIM ↑	CLIP ↑	Spatial ↑	Center ↓
pRF-Guide	10%	0.3197	0.3272	0.6182	0.5576	0.1212
No geometry	10%	0.3128	0.3094	0.5687	0.5373	0.1437
Native	10%	0.3151	0.3076	0.8327	0.5596	0.1171
pRF-Guide	25%	0.3187	0.3123	0.7238	0.5634	0.1128
No geometry	25%	0.3141	0.3103	0.7234	0.5494	0.1402
Native	25%	0.3227	0.3092	0.8749	0.5732	0.1066
pRF-Guide	50%	0.3246	0.3173	0.7959	0.5726	0.1076
No geometry	50%	0.3158	0.3154	0.7923	0.5595	0.1319
Native	50%		Not evaluated			
pRF-Guide	100%	0.3316	0.3195	0.8742	0.5802	0.1042
No geometry	100%	0.3180	0.3122	0.8709	0.5661	0.1267
Native	100%	0.3420	0.3357	0.9273	0.5884	0.1006

The four budgets contain nested sets of 795, 1,989, 3,979 and 7,959 fitting identities. Normalization, quality, target scales and the low-level mapper use full T. The 120-epoch schedules allow 3,000, 7,560, 15,000 and 29,880 optimizer steps, respectively. Actual successful updates and validation-selected checkpoint steps are provided in the accompanying metadata. These comparisons do not hold optimization steps fixed or establish whole-system data efficiency.

Historical quality is the geometric mean of normalized average ranks of pRF fit R^2 and released noise-ceiling SNR. The ROI factor is 1 for labels 1–7 and 0.5 otherwise. Equation 1 gives the quality-weighted relative log bias. The parameter-matched attention condition sets that bias to zero. The null permutes footprint rows within ROI, hemisphere, and quality strata with permutation seed 2027, retaining neural features and quality at their original voxels, and retrains the model with the same budget. The null preserves a structured spatial prior while breaking its measured correspondence.

The shared voxel feature transform and patch readout follow the architecture in Appendix A, but the historical global token comes from mean-pooled voxel features and a separate projection. Its projector and six-layer diffusion prior use the same four semantic loss weights $(1, 30, 5, 1)$. AdamW

Table 13: Paired pRF-Guide minus reference differences at each executed high-level supervision budget. Each cell reports the mean difference and nominal 95% paired-image bootstrap interval (982 images, 10,000 draws); no multiplicity correction is applied.

Budget	Reference	Spatial difference	Center difference (°)	GridSSIM difference
10%	No geometry	$+0.0203$ $[+0.0182, +0.0224]$	-0.0226 $[-0.0288, -0.0163]$	$+0.0179$ $[+0.0140, +0.0218]$
10%	Native	-0.0020 $[-0.0038, -0.0001]$	$+0.0040$ $[-0.0011, +0.0091]$	$+0.0196$ $[+0.0163, +0.0230]$
25%	No geometry	$+0.0140$ $[+0.0121, +0.0158]$	-0.0274 $[-0.0331, -0.0219]$	$+0.0020$ $[-0.0015, +0.0054]$
25%	Native	-0.0098 $[-0.0114, -0.0081]$	$+0.0062$ $[+0.0019, +0.0106]$	$+0.0031$ $[-0.0001, +0.0062]$
50%	No geometry	$+0.0131$ $[+0.0114, +0.0149]$	-0.0244 $[-0.0296, -0.0192]$	$+0.0018$ $[-0.0013, +0.0050]$
100%	No geometry	$+0.0141$ $[+0.0124, +0.0158]$	-0.0225 $[-0.0273, -0.0177]$	$+0.0073$ $[+0.0045, +0.0102]$
100%	Native	-0.0082 $[-0.0098, -0.0067]$	$+0.0036$ $[-0.0005, +0.0078]$	-0.0162 $[-0.0187, -0.0135]$

uses peak learning rate 3×10^{-4} and weight decay 10^{-2}. Four GPUs with microbatch 8 give batch 32. The one-cycle schedule has a two-epoch warm-up and final division factor 1,000; gradients are clipped to norm 1. CLIP and the image generator remain frozen. The standalone supplement retains the full operational specification.

H.2 Low-level development experiments

The core high-level study shares one low-level mapper within each participant. Participant 1 uses the official MindEye mapper; participants 2, 5 and 7 use a width-4,096, four-block mapper trained for 120 epochs. Historical fixed-fusion development for participants 2 and 7 compares true, shuffled and geometry-free local branches at $\alpha = 0.5$, using two local training seeds while fixing the high-level model. These 20-epoch local fits and their 300-image evaluations reuse the checkpoint-selection validation set. Their architectures, objectives, per-seed results and fitting details are retained in the standalone supplement; they are distinct from the expanded Participant 1 experiment.

I Metrics and statistical interpretation

I.1 Generation and global image measures

All controlled core conditions use 16 candidates, 100 prior steps, 20 image-diffusion steps, guidance 3.5, and img2img strength 0.85, with shared image/candidate seed schedules. Selection maximizes cosine similarity between flattened brain-projector tokens and frozen CLIP hidden tokens of each candidate. Candidate-mean scores evaluate the pixelwise average of the 16 images; they are distinct from averaging the candidates' individual metric values.

PixCorr is flattened RGB Pearson correlation after resizing to 425 pixels by bilinear interpolation. Global SSIM uses grayscale at that size, Gaussian weighting with $\sigma = 1.5$, population covariance, and data range 1. AlexNet layer 2/5, InceptionV3 and CLIP report two-way identification: for each reconstruction, the paired target-feature correlation is compared with correlations to every other target in the evaluation pool. Strictly greater comparisons count as successes and are averaged over distractors and images. EfficientNet-B1 and SwAV instead report paired feature correlation distance $1 - \rho$, where lower is better. The supplement specifies network layers, input sizes and normalization.

I.2 Spatial and local image measures

SwAV ResNet-50 layer-3 maps are extracted from 224-pixel inputs. Spatial cosine averages cosine similarity at corresponding map locations. With feature energy $e_u = \|F_u\|_2$ and cell coordinates p_u in degrees, the energy centroid is $\mu^F = \sum_u e_u p_u / \sum_u e_u$; centroid error is the Euclidean distance

between target and reconstruction centroids. Energy-region IoU compares the largest 20% of feature-energy cells. These are properties of generated-image feature maps, not direct measurements of attention or anatomical localization.

COCO boxes are mapped into the NSD crop. For original dimensions (W, H) and normalized crop margins (t, b, l, r), coordinates transform as $x' = \mathrm{clip}((x/W - l)/(1 - l - r), 0, 1)$ and $y' = \mathrm{clip}((y/H - t)/(1 - t - b), 0, 1)$. Pixel metrics retain boxes at least 12 target pixels in both dimensions and resize corresponding crops to 64×64 by bicubic interpolation. Local PixCorr uses RGB; historical local SSIM uses grayscale. Box feature cosine pools normalized SwAV vectors within mapped regions. All box scores are first averaged within each image. Undefined scores are omitted; paired analyses use jointly finite images. These COCO-box measures differ from the expanded GridSSIM metric defined in Appendix A.

I.3 Auxiliary image-to-fMRI evaluation

The ridge encoder predicts responses from frozen 2,048-dimensional SwAV average-pool features. Of the historical 8,559 training identities, the first 300 sorted IDs select the ridge penalty and 8,259 fit the model. Standardization uses the fitting set. The penalty grid is $0.01, 0.1, 1, 10, 100, 1000, 10000$, selected by response MSE. All four participants select 10,000. Mean per-image correlation compares predicted and measured responses across voxels; mean voxel correlation compares them across images. This fitted encoder is excluded from reconstruction training and selection. Real-image mean correlations for participants 1, 2, 5 and 7 are 0.3546/0.3681/0.4481/0.3169; the standalone supplement tabulates the corresponding reconstruction scores. These are references for this fitted evaluator, not biological ceilings. Its SwAV backbone also supplies the spatial image metrics, so these measures are related diagnostics.

I.4 Aggregation and uncertainty

We average image scores within runs, average three seeds within each participant, and weight the four participants equally. Main-table SDs are sample SDs across participants; Table 14 reports within-participant seed SDs. Seeds are repeated fits, not additional participants. True-minus-control contrasts are paired within participant and seed.

Historical image-bootstrap intervals use 10,000 paired image resamples. Saved sign-flip tests use 10,000 assignments and Holm adjustment across the eight spatial tests (four measures, two controls). The participant-level exact two-sided sign-flip test has only $2^4 = 16$ assignments and a minimum attainable p-value of 0.125. Accordingly, effect sizes and participant directions are the principal cross-participant summaries. The expanded Participant 1 intervals are nominal and unadjusted, as specified in the main text. Both forms of image interval condition on the evaluated participants, fitted models, and identification pool.

J Complete historical results

Tables 14–16 retain participant-level and full-metric outcomes. True geometry improves reconstructed-image spatial cosine over matched attention in all twelve participant–seed pairs. Against the null, cosine improves in ten pairs and centroid error in eleven; the corresponding participant means are favorable in all four participants. Mean COCO-box SSIM decreases relative to matched attention in all four participant means.

Candidate averaging changes the appearance/identification balance. For true geometry it increases PixCorr from 0.2933 to 0.3543 and global SSIM from 0.3193 to 0.4080, while reducing CLIP identification from 0.8720 to 0.7610. Candidate-mean images improve all eight aggregated global measures relative to both controls, but they are a different output policy. The analysis does not select a different output type for each metric.

K Low-level development and external systems

The standalone supplement retains the full validation table for participants 2 and 7. External comparisons align all 982 Participant 1 identities and use common evaluation code, while preserving

each system's native training and generation. The historical geometric-core row uses Participant 1 seed 42 and the original mapper. MindEye uses the official-baseline reconstruction set; MindEye2 uses its refined 40-session outputs; Brain-IT, SynMind and NeuroPictor use released reconstructions.

Table 14: Per-subject official-test results. Each entry is the mean ± sample SD over three high-level training seeds (42, 123, 2027), with the subject's low-level mapper held fixed.

Method	PixCorr ↑	SSIM ↑	CLIP ↑	Spatial ↑	Center ↓	Brain ↑
NSD Participant 1						
Learned attention	0.364 ± 0.001	0.318 ± 0.003	0.873 ± 0.006	0.575 ± 0.001	0.117 ± 0.003	0.326 ± 0.006
Shuffled pRF prior	0.372 ± 0.001	0.324 ± 0.002	0.871 ± 0.004	0.585 ± 0.001	0.101 ± 0.001	0.340 ± 0.001
pRF-Guide (core)	0.376 ± 0.002	0.326 ± 0.002	0.890 ± 0.003	0.590 ± 0.000	0.100 ± 0.001	0.341 ± 0.001
NSD Participant 2						
Learned attention	0.288 ± 0.004	0.314 ± 0.001	0.869 ± 0.003	0.567 ± 0.002	0.126 ± 0.005	0.345 ± 0.002
Shuffled pRF prior	0.294 ± 0.001	0.317 ± 0.003	0.864 ± 0.000	0.575 ± 0.000	0.107 ± 0.002	0.356 ± 0.001
pRF-Guide (core)	0.303 ± 0.001	0.321 ± 0.001	0.886 ± 0.006	0.582 ± 0.001	0.102 ± 0.001	0.359 ± 0.002
NSD Participant 5						
Learned attention	0.258 ± 0.002	0.316 ± 0.002	0.884 ± 0.002	0.562 ± 0.000	0.130 ± 0.001	0.413 ± 0.001
Shuffled pRF prior	0.266 ± 0.001	0.323 ± 0.005	0.880 ± 0.005	0.569 ± 0.001	0.118 ± 0.000	0.423 ± 0.000
pRF-Guide (core)	0.264 ± 0.000	0.323 ± 0.002	0.891 ± 0.005	0.572 ± 0.000	0.115 ± 0.002	0.425 ± 0.002
NSD Participant 7						
Learned attention	0.225 ± 0.001	0.303 ± 0.002	0.852 ± 0.002	0.560 ± 0.000	0.137 ± 0.001	0.297 ± 0.001
Shuffled pRF prior	0.224 ± 0.001	0.304 ± 0.001	0.820 ± 0.003	0.563 ± 0.001	0.123 ± 0.001	0.295 ± 0.005
pRF-Guide (core)	0.230 ± 0.001	0.307 ± 0.003	0.821 ± 0.006	0.564 ± 0.000	0.121 ± 0.001	0.298 ± 0.001

CLIP denotes two-way identification. Spatial is aligned SwAV layer-3 feature cosine; Center is feature-energy centroid error in degrees over an 8.4° field; Brain is mean per-image response correlation from the independent encoding evaluator.

Table 15: Complete global reconstruction metrics for the 36-run official-test study. Each entry averages training seeds 42, 123, and 2027 within each of participants 1, 2, 5 and 7, then reports the equal-weight subject mean ± sample SD across four subjects. Each configuration contains 982 test stimuli and 16 generated candidates per stimulus.

Metric	learned attention	Retrained null	True pRF
Selected reconstruction			
PixCorr ↑	0.284 ± 0.059	0.289 ± 0.062	0.293 ± 0.063
SSIM ↑	0.313 ± 0.007	0.317 ± 0.009	0.319 ± 0.008
AlexNet(2) ↑	0.873 ± 0.044	0.895 ± 0.045	0.907 ± 0.045
AlexNet(5) ↑	0.916 ± 0.024	0.934 ± 0.031	0.943 ± 0.030
InceptionV3 ↑	0.854 ± 0.013	0.838 ± 0.033	0.855 ± 0.041
CLIP ↑	0.869 ± 0.013	0.859 ± 0.027	0.872 ± 0.034
EfficientNet-B ↓	0.765 ± 0.011	0.773 ± 0.029	0.755 ± 0.041
SwAV ↓	0.452 ± 0.011	0.454 ± 0.024	0.441 ± 0.033
Pixelwise candidate-mean image			
PixCorr ↑	0.350 ± 0.072	0.351 ± 0.073	0.354 ± 0.073
SSIM ↑	0.406 ± 0.009	0.407 ± 0.010	0.408 ± 0.010
AlexNet(2) ↑	0.845 ± 0.058	0.853 ± 0.061	0.861 ± 0.060
AlexNet(5) ↑	0.840 ± 0.045	0.850 ± 0.049	0.859 ± 0.051
InceptionV3 ↑	0.675 ± 0.023	0.679 ± 0.030	0.689 ± 0.037
CLIP ↑	0.751 ± 0.015	0.746 ± 0.030	0.761 ± 0.036
EfficientNet-B ↓	0.933 ± 0.007	0.931 ± 0.013	0.925 ± 0.016
SwAV ↓	0.638 ± 0.005	0.634 ± 0.010	0.629 ± 0.014

The upper block scores the candidate returned by the fixed MindEye selector. The lower block scores the pixelwise average of the 16 candidates; it is not an average of candidate scores. AlexNet, InceptionV3, and CLIP are two-way identification scores. EfficientNet-B and SwAV are correlation distances. Arrows indicate the favorable direction.

Table 16: Spatial, local, and neural outcomes for selected reconstructions in the controlled 36-run official-test study. Method columns report mean ± sample SD across four subject means, each averaging three training seeds. The last two columns give the mean within-subject difference, true pRF minus learned attention (Δ_{attn}) or the shuffled prior (Δ_{null}); negative differences favor true pRF for center error and response MSE.

Metric	learned attention	Retrained null	True pRF	Δ_{attn}	Δ_{null}
Spatial structure					
Spatial cosine ↑	0.566 ± 0.007	0.573 ± 0.009	0.577 ± 0.012	$+0.0109$	$+0.0039$
Center error (deg) ↓	0.128 ± 0.008	0.112 ± 0.010	0.109 ± 0.010	-0.0181	-0.0027
Top-20% energy IoU ↑	0.230 ± 0.013	0.256 ± 0.017	0.262 ± 0.020	$+0.0322$	$+0.0060$
Box feature cosine ↑	0.740 ± 0.005	0.748 ± 0.008	0.752 ± 0.010	$+0.0119$	$+0.0037$
Local appearance					
Local PixCorr ↑	0.140 ± 0.032	0.143 ± 0.033	0.145 ± 0.033	$+0.0044$	$+0.0020$
Local SSIM ↑	0.106 ± 0.002	0.103 ± 0.003	0.103 ± 0.003	-0.0026	$+0.0005$
Independent neural evaluation					
Image response corr. ↑	0.345 ± 0.049	0.354 ± 0.053	0.356 ± 0.053	$+0.0103$	$+0.0019$
Response MSE ↓	0.402 ± 0.009	0.398 ± 0.009	0.398 ± 0.008	-0.0039	-0.0003
Mean voxel corr. ↑	0.293 ± 0.049	0.301 ± 0.052	0.302 ± 0.052	$+0.0081$	$+0.0007$
Median voxel corr. ↑	0.283 ± 0.058	0.293 ± 0.062	0.294 ± 0.063	$+0.0107$	$+0.0010$

Global spatial and neural scores use all 982 test images per configuration. Center error measures displacement of SwAV feature-energy centroids within an 8.4° field. Local scores average valid COCO-box regions after matching the NSD crop. Voxel correlations are summarized within each subject before aggregation across subjects; voxels and training seeds are not additional subjects.