DRAFT: DECODING AND RECONSTRUCTING 3D ROOM LAYOUTS FROM FMRI AND CAMERA TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

While functional magnetic resonance imaging (fMRI) decoders reconstruct images and 3D objects, recovering the metric geometry of rooms from a moving viewpoint remains an open challenge. We introduce, to our knowledge for the first time, the task of reconstructing 3D layouts of held-out scenes using only the known camera trajectories and the fMRI signals of participants watching indoor walkthroughs, without any visual frames. The task is hard because of the temporal mismatch between rapid visual changes and blurred fMRI signals. Our method, DRAFT (**D**ecoding and **R**econstructing 3D room l**A**youts from **f**MRI and **T**rajectories), extracts a compact layout representation from the early visual cortex and fuses this neural code with the geometric free space derived from camera motion. To evaluate it, we propose layout r, the Pearson correlation between predicted and ray-cast log depths. On ten held-out scenes (20 participants), DRAFT reaches a layout r of 0.480, outperforming the camera path alone (0.446) and seven fMRI decoders (at most 0.429). DRAFT also works as a plug-in that improves all seven decoders without retraining, raising their layout r from 0.171–0.429 to 0.487–0.493. Brain activity and camera motion can thus recover the metric layout of unseen rooms, a first step toward reconstructing the 3D spaces people see.

1 INTRODUCTION

Reconstructing a room needs more than a plausible picture. Walls must sit at the right distances, and views taken from different places must describe the same space. Decoders of functional magnetic resonance imaging (fMRI) now reconstruct images and videos (Ozcelik & VanRullen, 2023; Scotti et al., 2024; Chen et al., 2023b), depth images (Gaziv & Irani, 2021; Xia et al., 2024) and isolated 3D objects (Gao et al., 2024), yet to our knowledge none has recovered metric room geometry as the viewpoint moves. We ask whether fMRI responses to indoor walkthrough videos support layout and surface reconstruction in scenes the model has never seen.

This is hard for three reasons. The viewpoint changes every second, the blood-oxygen-level-dependent (BOLD) signal of fMRI blurs several seconds together, and recordings are few. So we do not decode depth maps directly. We decode a small layout code, 22 numbers per second that describe wall distances, floor extent, openings and depth range. Ridge regressions, the readout, predict it from early visual cortex (EVC) and camera motion. EVC responses map the image point by point, so they carry image structure that correlates with depth. Scene-selective regions also respond to room layout (Section 2), but in our data EVC predicted the code better (Section 5.2). The camera path also helps. Every visited position is free space, just as steps through a dark room show where the walls are not. We use it as a learned prior, not a bound on wall distance.

Our method DRAFT (Figure 2) combines both and, like an architect's draft, fixes the coarse layout first and the details later. It renders the code as coarse depth, and a small generator trained on real readout errors adds detail to eight consecutive views. DRAFT then replaces the broad structure (low spatial frequencies) of the views with a 3×3 DRAFT layout, predicted from the code and the camera path, and places the views in 3D at the known poses.

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.

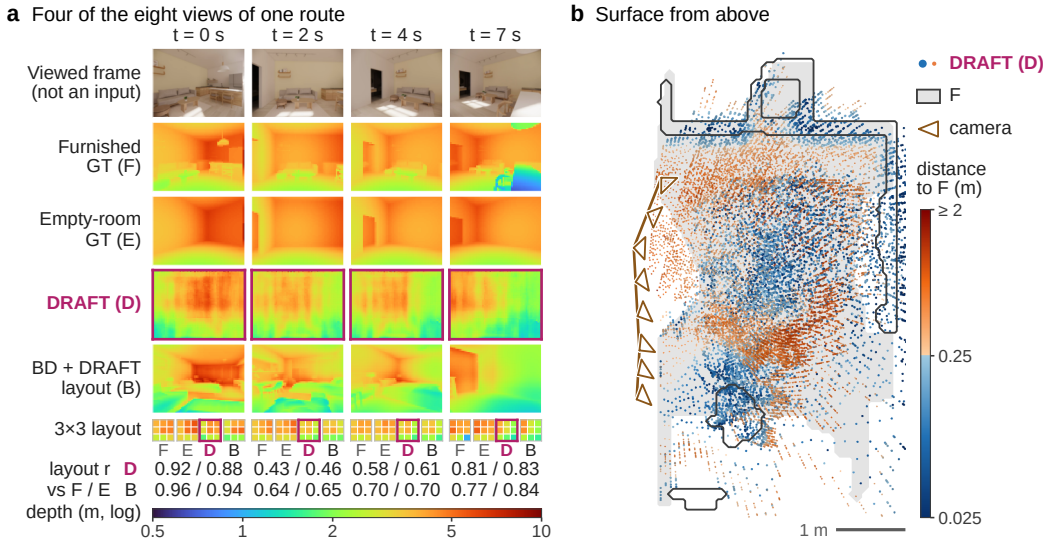


Figure 1: **DRAFT recovers where walls and floor are on a held-out route.** **(a)** Four views of scene 4: frame (not an input), furnished (F) and empty-room (E) ground truth, DRAFT (D), BrainDiffuser (BD) detail with the DRAFT layout (B), and layout r of D and B against F and E. **(b)** Merged surface of D from above, blue within 25 cm of F. Hand-picked route, ranked 60th in layout r and 94th in F-score of 324 routes. Rule-chosen routes: Figure 7 and Section K.

Across ten held-out scenes, the DRAFT layout's cell depths correlate with the true ones at 0.480 (layout r, Section 5.1). This is above the best of seven adapted fMRI decoders (0.429), the fixed template (median training-scene depth, 0.324) and the generator alone (0.303). Most of this comes from the camera path (0.446), but the brain adds to it. Adding the readout to the camera path raises layout r in all ten scenes, and brain-input controls lower it in all ten (Section 5.2).

The 3D surfaces tell a weaker story (Section 5.4). Because ridge regressions shrink the depth range, DRAFT ends with a range calibration, whose factor we fixed after earlier results on the same scenes. Even so, unmodified MindEye and Brain-Diffuser have slightly better mean surface scores, and so does the generator alone with a milder calibration. Guided by the DRAFT layout, these two image decoders reach the best fMRI-based mean surface scores, within noise of their unmodified versions.

We contribute a readout of a layout code from EVC and camera motion, and the DRAFT layout, which fuses this code with the camera path and raises the layout r of all seven tested decoders without retraining (five post hoc). We also contribute a held-out-scene test of room geometry from fMRI, with ray-cast ground truth, fMRI-free and wrong-scene references and brain-input controls.

2 Related Work

Image, depth and video reconstruction from fMRI. fMRI decoding reconstructs movies (Nishimoto et al., 2011) and images (Shen et al., 2019) through learned feature mappings (Kay et al., 2008; Naselaris et al., 2011). The Natural Scenes Dataset (NSD) (Allen et al., 2022) supports diffusion-based image decoding (Takagi & Nishimoto, 2023a; Ozcelik & VanRullen, 2023; Scotti et al., 2023; 2024). MinD-Vis adds brain pretraining and MinD-Video extends it to video (Chen et al., 2023a;b). Gaziv & Irani (2021) reconstruct depth and color images with monocular-depth supervision, and DREAM decodes depth, color and semantics for image reconstruction (Xia et al., 2024). We instead test metric layout in held-out scenes against ray-cast geometry.

3D decoding and scene-selective cortex. MinD-3D reconstructs object meshes from fMRI of rotating objects (Gao et al., 2024), and Neuro-3D reconstructs colored object point clouds from electroencephalography (EEG) (Guo et al., 2025). Closest to our setting, Zeng & Cai (2024) decode the depth structure of first-person driving videos in a held-out virtual town by matching decoded embeddings to those of the town's own scenes. DRAFT instead predicts metric depth without the held-out

scene's geometry and lifts it to 3D surfaces. Scene-selective regions respond to spatial layout (Epstein & Kanwisher, 1998; Dilks et al., 2013; Henriksson et al., 2019), boundaries (Park et al., 2011; Julian et al., 2016), navigational affordances (Bonner & Epstein, 2017) and 3D surface configuration (Lescroart & Gallant, 2019), and support the cognitive map (Epstein et al., 2017). For static photographs, Gabor-wavelet features predict their responses better than explicit surface models (Shafer-Skelton et al., 2026), and our readout also cannot separate depth from correlated image features.

Depth estimation and generative priors. Monocular depth models (Eigen et al., 2014; Ranftl et al., 2022; 2021; Yang et al., 2024a;b; Bochkovskii et al., 2025; Ke et al., 2024) now learn from large mixed datasets and diffusion priors. Replacing low spatial frequencies resembles low-frequency guidance of diffusion models (Choi et al., 2021) and spatial control of pretrained generators (Zhang et al., 2023). Our replacement acts on output depth without training, so any decoder can supply the detail. The camera path as a free-space prior relates to occupancy mapping (Elfes, 1989) and trajectory-based free-space prediction (Wijmans et al., 2023).

3 Task, Data and Evaluation

Task. A participant watches a first-person walkthrough video of a furnished indoor scene during fMRI. We split each video into *routes* of eight consecutive one-second views. For each route, the model predicts eight depth maps and the 3D surface they form. Its inputs are the brain responses and the camera path (position, heading and pitch in each second), never the video frames.

Data. Twenty participants, 6 to 19 per scene, watched 80-second videos of ten furnished indoor scenes (Table 4). Each scene was rendered in up to four interior styles that share the floor plan and a similar camera path but differ in materials, colors and furniture. This gives 36 videos and 416 recordings (one participant watching one video). Each recording has 74 one-second samples, because the last six seconds are left for the response delay. Viewing was free, with no fixation task or eye tracking, and the same participants appear in training and held-out scenes. Ground-truth radial depth (distance along each camera ray) is ray-cast from the 3D scene models at the recorded poses (90° horizontal field of view, 48×64 pixels). Nine non-overlapping routes per video give 324 routes. Nuisance regression and z-scoring use whole scanner runs, which can mix training and held-out scenes but never a depth target (Section A.1).

Protocol. We use leave-one-scene-out cross-validation over the ten scenes. Every model behind a test prediction is fitted without any style, recording or route of the held-out scene. Models that take other models' predictions as input are fitted on out-of-fold predictions: for each training scene, models fitted without that scene predict it (Section A.2). At test time, predictions are averaged over the recordings of the held-out video and over another style at matching poses, without depth labels (Section 4.1). So results use pooled recordings, not single trials or single participants.

Comparisons. We adapt seven fMRI decoders to our folds. Five are published: image and depth decoding (Takagi & Nishimoto, 2023a;b), the DREAM depth branch (Xia et al., 2024), Brain-Diffuser (Ozcelik & VanRullen, 2023) and MindEye (Scotti et al., 2023). Two are simple ridge decoders: linear depth decoding and retrieval of CLIP (Contrastive Language-Image Pre-training) image embeddings (Radford et al., 2021), like the retrieval branch of MindEye. The published decoders keep their released components and original brain regions. Depth Anything V2 (DA V2) (Yang et al., 2024b), with its scale calibrated on the training scenes, turns decoded images into radial depth. The decoders may read scene-selective regions and use large pretrained models, but they see no camera input and decode each view separately. Trained on nine scenes, Brain-Diffuser and MindEye identify the viewed frame only weakly (Section C). We also use two fMRI-free references. The *fixed template* is the median training-scene depth, and the per-second median is this median taken separately at each video second (Section G).

Metrics. Our main metric, **layout** r, asks whether a prediction knows which parts of a view are near or far. It is the Pearson correlation between the predicted and true median log depths of the nine cells of a 3×3 grid over the view. **Row order** is the fraction of correctly ordered cell pairs in the same row, and **left/right** the fraction of views whose deeper half is identified. **Pixel MAE**

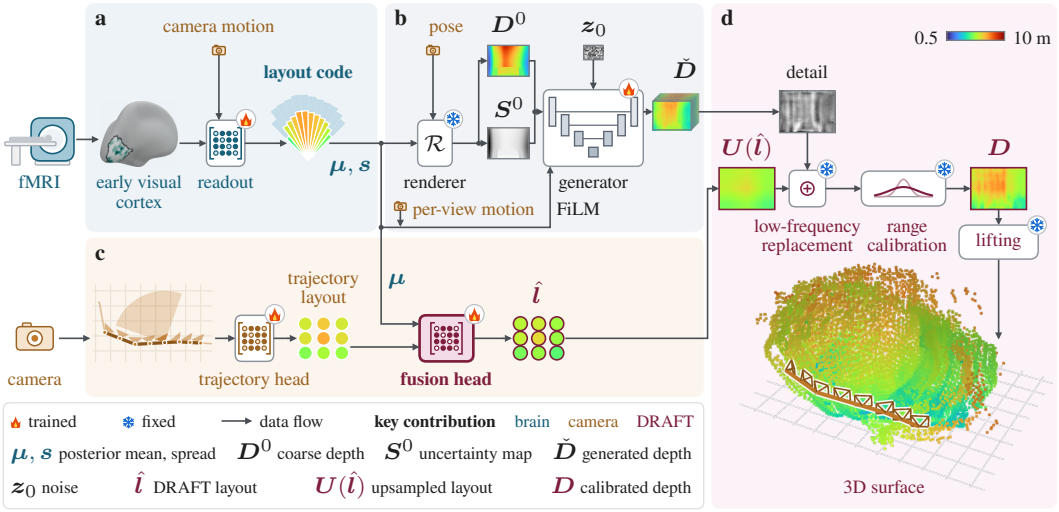


Figure 2: **Overview of DRAFT: the coarse layout is fixed before the detail.** (a) The readout maps EVC and camera motion to the layout code, which (b) a renderer and generator turn into eight depth maps. (c) Trajectory and fusion heads give the DRAFT layout, which (d) sets the broad structure of any depth map before range calibration and lifting to 3D. Real data of the Figure 1 route, except the noise $\boldsymbol{z}_0$.

(mean absolute error) is the depth error in meters over valid pixels. All four are averaged over the 324 routes (Section B.1). For surfaces, we back-project each view along its camera rays at the known pose. The eight views of a route then merge on 5 cm voxels into one point set, the surface, without extracting a mesh. We compare it with the ground-truth points visible from the same poses, not with the whole room mesh. The **F-score** (F@25cm) is the harmonic mean of **precision**, the fraction of predicted points within 25 cm of a true point, and **recall**, the fraction of true points within 25 cm of a predicted point. The **Chamfer distance** averages the nearest-point distances in both directions, in meters. Surface metrics and paired contrasts weight scenes equally, with 95% confidence intervals (CIs) from 20,000 paired scene-bootstrap resamples and the number of scenes that favor each method. Intervals are not corrected for multiple comparisons and can be too narrow, as scenes share participants and some scanner runs. So four main layout contrasts, named after the results were known, also get sign-flip p values with Holm correction. Some other contrasts get uncorrected p values. An interval that includes zero does not prove two methods equal (Section B).

4 Method

4.1 Readout of the layout code

Layout code. The code $\boldsymbol{y}(\boldsymbol{D}) \in \mathbb{R}^{22}$ summarizes one view's depth map $\boldsymbol{D}$ in log radial depth,

$$\boldsymbol{y}(\boldsymbol{D}) = \big[\, w_1, \ldots, w_{12},\ f_1, \ldots, f_4,\ o_1, \ldots, o_4,\ \operatorname{med}\log \boldsymbol{D},\ \operatorname{IQR}\log \boldsymbol{D} \,\big]. \tag{1}$$

Here w_k is the wall distance near the horizon in azimuth bin k, f_b the floor extent in vertical band b, o_q the share of far pixels in quadrant q and IQR the interquartile range (Section A.2).

Readout. Six ridge regressions map EVC responses and 12 camera-motion regressors to the code $\boldsymbol{y}$, standardized on the training scenes. They differ in voxel count, ridge penalty and response window, and their mean is the readout $\hat{\boldsymbol{y}}_t$ (Equation (5)). EVC comprises V1, V2, ventral V3 (V3v) and human V4 (hV4), whose most reliable voxels are reduced to 16 principal components. The regressors are a constant, speed and turning speed at lags of 0 to 4 s, and a linear trend. The regressions are fitted to participant-averaged responses of the nine training scenes and applied to each held-out recording.

Pooling across styles. One recording is noisy, and other styles of a scene often pass the same poses. So each second we average the readout of the evaluated recording with the participant-averaged readout of at most one other style, weighted by recording counts. Pooling never uses depth labels or pairs different seconds, but furniture may differ across styles (Section A.3). The pooled readout of each code entry j is its posterior mean μ_j. Because readout reliability varies, each entry also gets a posterior spread s_j, its out-of-fold residual standard deviation (Section A.2). Together they form a point estimate with an error bar, not a full Bayesian posterior.

4.2 Coarse depth and flow-matching generator

Renderer. A renderer turns the posterior mean and pose into coarse inverse depth $\boldsymbol{D}_t^0$. This map holds walls, eye height, floor extents and openings, matched to the code's median and IQR. A first-order approximation carries the posterior spreads into an uncertainty map. It follows only each pixel's dominant wall distance, not the renderer's full covariance (Section A.2).

Generator. Coarse depth has walls and floors but no detail, so a generator adds it, modeling the eight views of a route jointly. It learns the residual between the true inverse depth $\boldsymbol{Q}_{\mathrm{GT}}$ and $\boldsymbol{D}^0$ with rectified flow (Liu et al., 2023), a form of flow matching (Lipman et al., 2023). A penalty asks the generated depth to agree with the decoded layout code, most strongly where the code is reliable,

$$\mathcal{L} = \mathbb{E}\left\|v_\theta(\boldsymbol{z}_\tau, \tau, \boldsymbol{c}) - (\boldsymbol{z}_1 - \boldsymbol{z}_0)\right\|^2 + \lambda \ \mathrm{mean}\left[\widetilde{\boldsymbol{w}}\left(\mathcal{A}_i(\hat{\boldsymbol{Q}}_1) - \boldsymbol{\mu}\right)^2\right], \qquad \boldsymbol{z}_\tau = (1-\tau)\boldsymbol{z}_0 + \tau \boldsymbol{z}_1. \quad (2)$$

Here $\tau \in [0, 1]$ is the flow time, v_θ the velocity of a network with weights θ, $\boldsymbol{z}_1 = (\boldsymbol{Q}_{\mathrm{GT}} - \boldsymbol{D}^0)/\rho$ the residual scaled by a training-scene constant ρ, $\boldsymbol{z}_0$ Gaussian noise and $\boldsymbol{c}$ the conditions (coarse depth, uncertainty map, posterior and motion). The penalty acts on the predicted depth $\hat{\boldsymbol{Q}}_1 = \boldsymbol{D}^0 + \rho\left(\boldsymbol{z}_\tau + (1-\tau)v_\theta\right)$. A fixed affine inverse of the renderer, $\mathcal{A}_i$, returns the code implied by a depth map, standardized like the posterior mean of the sampled recording i. The weights $\widetilde{\boldsymbol{w}}$ are the normalized inverse variances s_j^{-2}, and $\lambda = 0.1$ (Section A.2). Without the penalty, the generator alone has a higher layout r (0.418 against 0.301, ten-scene means), but DRAFT has worse surfaces (Section H.1). The network is a 3D U-Net (Çiçek et al., 2016) over space and views, conditioned by feature-wise linear modulation (FiLM) (Perez et al., 2018). Training pairs each route's ray-cast depth with the out-of-fold posterior of a random recording of its video, so the conditions contain real readout errors. At test time, $\check{\boldsymbol{D}}_t$ inverts the mean inverse-depth sample over the video's recordings.

Loss correction. An earlier implementation compared the two sides of the penalty in different coordinates, so all main results use ten generators retrained with the corrected loss (Section H). The correction lowers the layout r of the generator alone from 0.369 to 0.303. The DRAFT layout barely changes (0.486 to 0.480), so part of the fusion gain reflects a weaker generator.

4.3 Trajectory head, fusion and range calibration

Trajectory head. The camera path also carries layout information. In each of 12 directions, a free-space feature is the largest distance from the camera to a position visited within a few seconds, in that direction (Equation (7)). The trajectory head, a ridge regression on these and other path features, predicts a 3×3 log-depth layout, the trajectory layout (Equation (3)).

Fusion. The code and the trajectory layout $\boldsymbol{l}_t^{\mathrm{T}}$ make different errors, so a second ridge regression, the fusion head, combines them as a stacking model (Wolpert, 1992),

$$\hat{\boldsymbol{l}}_t = \boldsymbol{W}_{\mathrm{F}}^\top\left[\boldsymbol{\mu}_t;\, \boldsymbol{l}_t^{\mathrm{T}}\right] + \boldsymbol{b}_{\mathrm{F}}, \qquad \boldsymbol{l}_t^{\mathrm{T}} = \boldsymbol{W}_{\mathrm{T}}^\top \boldsymbol{\phi}_t + \boldsymbol{b}_{\mathrm{T}}, \quad (3)$$

where $\boldsymbol{\phi}_t$ holds the 111 path features of the trajectory head. The fusion head is fitted on out-of-fold inputs, and inner cross-validation on the same inputs, without further nesting, chooses its penalty (Section A.2). Its output $\hat{\boldsymbol{l}}_t \in \mathbb{R}^9$ is the DRAFT layout, which sets the low spatial frequencies of the log depth. The generator supplies the detail. This low-frequency replacement has no fitted parameters, so it applies to any other decoder's depth as well (Section 5.3),

$$\log \boldsymbol{D}_t = G_\sigma * \boldsymbol{U}(\hat{\boldsymbol{l}}_t) + \left(\log \check{\boldsymbol{D}}_t - G_\sigma * \log \check{\boldsymbol{D}}_t\right), \qquad \sigma = 4 \text{ pixels}, \quad (4)$$

where $\boldsymbol{U}$ is bilinear upsampling to 48×64 and $G_\sigma *$ Gaussian low-pass filtering. Layout r scores these low frequencies, so it depends mainly on the DRAFT layout, not on the detail (Section G).

Table 1: **Main results: fusion gives the best layout at no measurable cost in surface accuracy.** Each layout replaces the broad structure of one generator's depth. C: EVC, M: camera motion (12 readout regressors, 72 trajectory-head features), T: free space. Row: row order, L/R: left/right, bold: best, underline: second. Layout, MAE: 324-route means, others: ten-scene means. Last rows: DRAFT minus generator alone (*: 95% paired scene-bootstrap CI excludes zero).

		Layout			Pixel	Surface	
Layout from	Inputs	$r\uparrow$	Row$\uparrow$	L/R$\uparrow$	MAE (m)$\downarrow$	F@25cm$\uparrow$	Chamfer (m)$\downarrow$
Generator alone	none	0.303	0.626	0.602	1.507	0.478	0.449
Readout, 3×3 target	C, M	0.440	0.656	0.621	<u>1.176</u>	**0.485**	<u>0.417</u>
Trajectory head	M	0.390	0.665	0.632	1.264	0.447	0.443
Trajectory head	M, T	<u>0.446</u>	<u>0.688</u>	<u>0.658</u>	1.252	0.465	0.427
Fusion head (DRAFT)	C, M, T	**0.480**	**0.696**	**0.668**	**1.153**	<u>0.484</u>	**0.405**
DRAFT minus generator alone		+0.172*	+0.066*	+0.065*	−0.356*	+0.006	−0.044*
Scenes with DRAFT better		10/10	10/10	8/10	10/10	4/10	8/10

Table 2: **Comparison with fMRI decoders: every decoder scores best with the DRAFT layout, and none matches it given only the camera path.** Alone: unmodified. + path, + DRAFT: with the trajectory layout (camera path only) or the DRAFT layout, post hoc except for linear depth and Brain-Diffuser. Averaging as in Table 1. Bold: row best. *: 95% CI of the gain over alone excludes zero (Table 5). Unmodified MindEye and Brain-Diffuser keep slightly higher F-scores than DRAFT. Decoders: 418 recordings, DRAFT: 416. ‡DA V2 on the frames. †Our adaptation (Section C.1).

	Layout $r\uparrow$			F@25cm$\uparrow$			Chamfer (m)$\downarrow$		
Method	alone	+ path	+ DRAFT	alone	+ path	+ DRAFT	alone	+ path	+ DRAFT
Linear depth	0.429	0.454	**0.490***	0.393	0.440	**0.455***	0.508	0.454	**0.436***
CLIP retrieval	0.261	0.451	**0.488***	0.455	0.473	**0.493***	0.421	0.406	**0.383***
Image decoding	0.171	0.450	**0.488***	0.442	0.456	**0.474***	0.446	0.426	**0.403***
Depth decoding	0.322	0.455	**0.491***	0.317	0.435	**0.449***	0.727	0.462	**0.444***
DREAM depth branch†	0.428	0.450	**0.488***	0.368	0.435	**0.449***	0.561	0.464	**0.445***
Brain-Diffuser	0.324	0.458	**0.493***	0.492	0.476	**0.499**	0.389	0.400	**0.377**
MindEye	0.270	0.451	**0.487***	0.494	0.476	**0.499**	0.381	0.406	**0.376**
*Best with the DRAFT layout (with *)*			7/7 (7)			7/7 (5)			7/7 (5)
DRAFT (own generator)	–	–	0.480	–	–	0.484	–	–	0.405
Camera path alone	–	0.446	–	–	0.465	–	–	0.427	–
Fixed template	0.324	–	–	0.391	–	–	0.524	–	–
Video frames + DA V2‡	0.865	–	–	0.742	–	–	0.196	–	–

Range calibration. Ridge regressions shrink predictions toward the mean, so the cell log depths of the DRAFT layout have only 0.49 of the true spread (generator 0.78). Before lifting, we therefore map $\log \boldsymbol{D}_t$ to $m + a(\log \boldsymbol{D}_t - m)$, with m the video's mean predicted log depth, which uses no test depth. This map keeps layout r, row order and left/right except where depth is clipped. Inner validation on each fold's training scenes selects $a = 2$ in every fold by Chamfer distance, or 2 or 2.5 by F-score (Section G). We added this step and its selection after earlier results on these scenes.

5 Experiments

5.1 Layout accuracy

The DRAFT layout is more accurate than every fMRI decoder that does not use it, even with the camera path. Its layout r is higher than that of linear depth in seven scenes and of MindEye and Brain-Diffuser in all ten (no sign-flip test, Table 5). Unlike these decoders, the fusion head is fitted to the cell depths that layout r scores. Given the camera path through the trajectory layout, each decoder still trails DRAFT by 0.024 to 0.032 (ten-scene means). It trails in eight to ten scenes (sign-flip $p \leq 0.010$), and this comparison is post hoc for five decoders (Table 5).

The gain is more than a good guess about indoor rooms. Indoor views share a common shape, floor near and walls farther away, so a template gets part of the layout right. The DRAFT layout

Table 3: **Layout r gains of DRAFT hold in nine or ten scenes, and surface gains are smaller.** Ten-scene means of paired differences (positive: first better), 95% paired scene-bootstrap CI, scenes favoring the first. r: layout r, BD: Brain-Diffuser, M, T: as in Table 1, cosine: similarity of centered codes. MAE, Chamfer in m. Sign-flip p: [a]0.002, [b]0.004, [c]0.012, [d]0.025, others untested.

Comparison	Metric	Gain [95% interval]	Scenes
DRAFT vs generator	r	+0.172 [0.135, 0.210]	10[a]
	Row	+0.066 [0.040, 0.094]	10
	L/R	+0.065 [0.028, 0.101]	8
	MAE	+0.356 [0.304, 0.408]	10
vs fixed template	r	+0.144 [0.084, 0.204]	9[b]
vs per-second median	r	+0.225 [0.149, 0.298]	10[a]
BD vs fixed template	r	−0.008 [−0.055, 0.039]	5
MindEye vs fixed template	r	−0.062 [−0.098, −0.027]	2
Generator vs fixed template	r	−0.028 [−0.068, 0.009]	2
Readout vs trajectory (M)	r	+0.050 [0.015, 0.092]	8[d]
Trajectory (M, T) vs (M)	r	+0.050 [0.022, 0.076]	8[c]
DRAFT vs trajectory head	r	+0.035 [0.025, 0.046]	10[a]
	MAE	+0.103 [0.068, 0.133]	9
on BD depth	r	+0.036 [0.025, 0.048]	10[a]
on linear depth	r	+0.037 [0.026, 0.048]	10[a]

Comparison	Metric	Gain [95% interval]	Scenes
Readout vs motion alone	r	+0.098 [0.053, 0.151]	9[b]
DRAFT vs readout	Row	+0.040 [0.023, 0.057]	10
EVC vs motion alone	Cosine	+0.219 [0.150, 0.290]	10[a]
EVC vs scene-selective	Cosine	+0.160 [0.102, 0.226]	9
DRAFT vs replaced EVC	r	+0.037 [0.025, 0.051]	10[a]
	F-score	+0.036 [0.024, 0.050]	10[a]
vs shifted BOLD	F-score	+0.044 [0.029, 0.060]	10[a]
DRAFT vs generator	Chamfer	+0.044 [0.018, 0.069]	8
	Precision	+0.055 [0.041, 0.071]	10
	F-score	+0.006 [−0.013, 0.026]	4
DRAFT vs no detail	F-score	+0.042 [0.036, 0.047]	10
	Chamfer	+0.049 [0.040, 0.057]	10
DRAFT vs wrong scene	F-score	+0.069 [0.042, 0.098]	10
minus BD's gain	F-score	+0.010 [−0.019, 0.042]	5

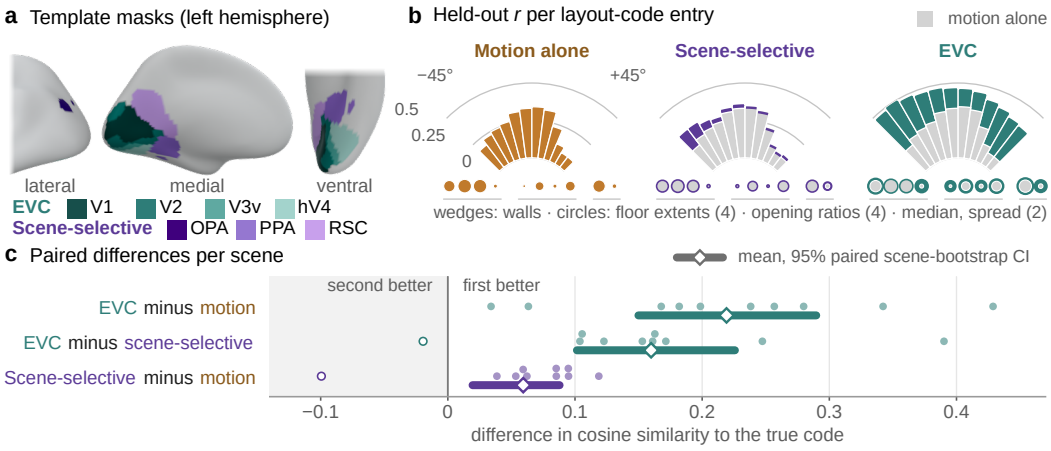


Figure 3: **EVC predicts the layout code best of the tested regions.** **(a)** Template masks, not individual localizers. **(b)** Held-out r per layout-code entry (ten-scene mean), from motion alone or with 1,000 voxels of one mask. **(c)** Per-scene differences in cosine similarity to the true code.

beats the fixed template in nine scenes and the per-second median (layout r 0.241) in all ten. Brain-Diffuser, MindEye and the generator alone do not beat the fixed template (Figure 4e).

5.2 Camera path and brain contributions

The camera path gives most of the layout, and the brain adds to it in every scene. From the camera path alone, the trajectory head reaches a layout r of 0.446 (Table 1). To score the readout by layout r directly, we also fit it to the 3×3 cell depths. Adding EVC to its 12 camera-motion regressors raises layout r from 0.338 to 0.440 in nine scenes (Table 3). Fusing the code with the trajectory layout raises layout r by 0.035 in all ten scenes and lowers pixel MAE in nine.

Two brain-input controls separate the brain's gain from camera motion. Because the readout also uses camera-motion regressors, the fusion gain does not isolate the brain. Both controls keep these regressors and lower the layout r and F-score of DRAFT in all ten scenes. Replacing the EVC features (16 principal-component scores) with their training-scene mean (replaced EVC) lowers them by 0.037 and 0.036 (Section D.3). Circularly shifting each recording's BOLD series by 37 of 80 seconds (shifted BOLD) and refitting the readout lowers them by 0.043 and 0.044 (one fixed shift, not a permutation test, Section D.4). At the readout level, the EVC gain peaks at a lag of 3 to 4 s and beats all 20 circular shifts (Section D.2). Post hoc, the brain's gain persists in single

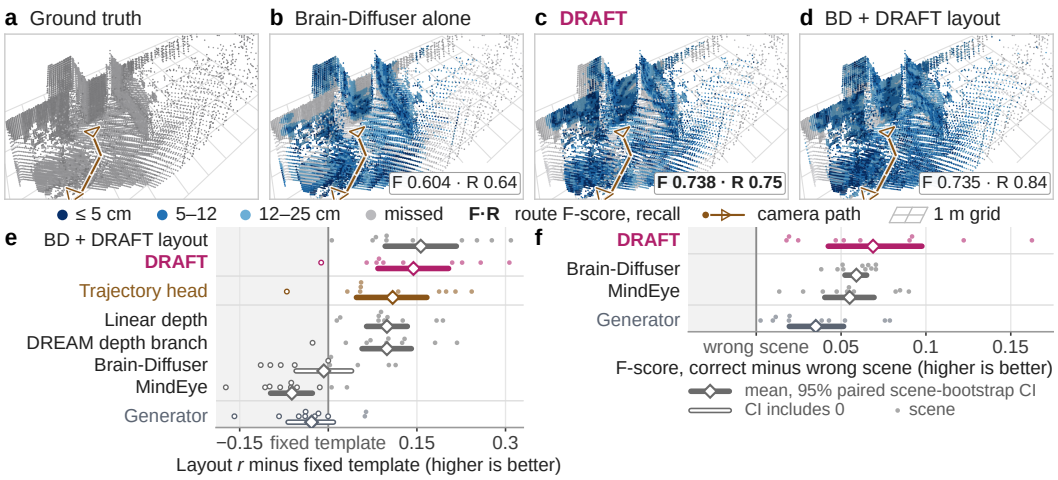


Figure 4: **Surface scores separate methods less than layout scores do.** (a to d) Display route (scene 6), one of the top five for DRAFT among routes where DRAFT and BD + DRAFT layout beat unmodified Brain-Diffuser (BD) in F-score. b to d: true points below 2.4 m by distance to the nearest prediction (F: F-score, R: recall, Figure 6). On rule-chosen routes, DRAFT has the lowest F-score of the generator, DRAFT and BD + DRAFT layout in seven of ten (Table 9). (e) Layout r gain over the fixed template. (f) Post hoc F-score gain over a wrong scene.

recordings (0.012, eight scenes) and in all 300 tested pipeline settings. It is no larger where the path covers less of the view, and more recordings would add only about 0.002 (Section I).

EVC predicts the code best among the tested regions. With 1,000 voxels per mask and the camera-motion regressors, EVC predicts the code better than motion alone in all ten scenes. It also beats the scene-selective occipital place area (OPA), parahippocampal place area (PPA) and retrosplenial complex (RSC) together in nine (Figure 3, Section D.2). Without functional localizers or dorsal motion areas, this only ranks the masks and shows no biological cause.

5.3 Guiding other decoders and agreement across views

Other decoders inherit the DRAFT layout. All seven decoders score best with it in all three metrics of Table 2, with a 95% CI excluding zero in 17 of 21 cases, all but the surfaces of Brain-Diffuser and MindEye. For Brain-Diffuser, it raises layout r from 0.324 to 0.493 in all ten scenes, the highest fMRI-based value. The other six reach about 0.49 (five post hoc), above their versions with the trajectory layout in all ten scenes (Section C.2). Layout r then mostly scores the DRAFT layout (Section 4.3), so surfaces are the real test.

The DRAFT layout also makes the views agree. The DRAFT layout lowers the change in 3D error between views 1 s apart for every scored depth source. Yet this measure also rewards views that are wrong in the same way (Table 11 and Section F.2). DRAFT views agree more than Brain-Diffuser's with the DRAFT layout (mean change 0.63 against 0.81 m, all ten scenes, Section F.2). But linear depth with the same layout, also decoded view by view, agrees as well or better (0.58 m).

5.4 Eight-view 3D surfaces

After range calibration, the DRAFT layout costs no measurable surface accuracy. Calibration brings the cell log depths of DRAFT to 0.97 of the true spread. DRAFT has a lower Chamfer distance than the generator alone (eight scenes), a higher precision (all ten) and a similar F-score (Table 3). But its Chamfer distance is not lower than the generator's at its own selected factor (1.5, Table 15). The readout with a 3×3 target, without free space, reaches a similar F-score (0.485). Without calibration, fusion lowers the F-score of the generator alone from 0.477 to 0.416 and raises its Chamfer distance by 0.065 m in all ten scenes. The DRAFT layout also lowers the uncalibrated F-score of the three decoders with the best unmodified surfaces, because ridge heads shrink the depth range (Table 13 and Section C.2).

The calibration factor was fixed after earlier results on these scenes. Selecting the factor by F-score instead of Chamfer distance picks 2 or 2.5 and lowers the held-out F-score of DRAFT to 0.477. At their Chamfer-selected factors, DRAFT, the generator alone, Brain-Diffuser and MindEye differ in F-score by at most 0.016, the generator alone slightly ahead, within noise (Table 15).

Unmodified MindEye and Brain-Diffuser reach slightly higher mean F-scores than DRAFT (0.494 and 0.492 against 0.484) and lower Chamfer distances (Table 2). Only the lower Chamfer distance of MindEye holds beyond noise (nine scenes, Section G). Their detail comes from large pretrained image models, ours from a small generator trained on nine scenes. Removing our detail lowers the F-score by 0.042 in all ten scenes (Table 3), but this detail is mostly one pattern shared by all scenes and correlates with the true detail at only 0.026 (Section G). So scene-specific geometry comes mainly from the DRAFT layout. Against decoders with the trajectory layout, DRAFT is within noise of CLIP retrieval, Brain-Diffuser and MindEye in F-score. It is ahead of the other four, each in at least eight scenes (post hoc sign-flip tests, Table 5). With the DRAFT layout, Brain-Diffuser's detail raises recall over DRAFT and slightly lowers precision, in all ten scenes. This gives the best fMRI-based mean F-score (0.499), within noise of both unmodified decoders (Section G). MindEye with the DRAFT layout ties it post hoc and has the lowest Chamfer distance (0.376 m).

Surface scores separate the methods little, for two reasons. First, even a wrong scene scores well. Depth predicted for another scene and lifted at the evaluated poses (post hoc) keeps most of the F-score of DRAFT, Brain-Diffuser and MindEye. The correct scene adds only 0.055 to 0.069, in all ten scenes each (Table 12, Figure 4f). Subtracting this wrong-scene score does not favor DRAFT over MindEye (post hoc, Section I). Second, the scores reward depth range. DA V2 gives decoded images nearly the true range, 0.84 of the true spread against 0.46 to 0.54 for our ridge layouts (Section C.1). This raises recall even for a wrong layout.

Once depth scale is removed, DRAFT has lower depth error than the unmodified image decoders, but this is a low bar. Scored without range calibration for every method, DRAFT has a lower scale-invariant depth error (SI-log, Eigen et al., 2014) than the trajectory head, the generator, Brain-Diffuser and MindEye in all ten scenes (sign-flip $p = 0.002$). After post hoc median alignment of each view to the true depth, its F-score beats these four in nine or ten scenes each. Yet Brain-Diffuser and MindEye have a higher SI-log than the fixed template (57.3 and 59.5 against 52.8, and 49.6 for DRAFT). Linear depth ties DRAFT, and Brain-Diffuser with the DRAFT layout nearly ties it in SI-log and beats it after alignment. The fixed template, Brain-Diffuser and MindEye also have more pixels within a factor of 1.25 of the true depth (Section F.1 and Table 10).

6 Discussion and Conclusion

To our knowledge, DRAFT is the first method to decode from fMRI the metric layout of rooms seen from a moving viewpoint and to test it against ray-cast ground truth in held-out scenes. It reads a layout code each second from EVC and camera motion, fuses it with the camera path and turns it into depth maps and 3D surfaces. In ten held-out scenes, its layout r is 0.480, above seven adapted fMRI decoders (at most 0.429) and the camera path alone (0.446), and brain-input controls lower it in every scene. Used as a plug-in prior without retraining, the DRAFT layout raises the layout r of all seven decoders to about 0.49 (five post hoc) and gives Brain-Diffuser and MindEye the best fMRI-based mean surface scores. Its fMRI-free and wrong-scene references and brain-input controls separate what the camera path, the brain and pretrained image models contribute.

Limitations. DRAFT needs known camera paths and pooled recordings, was developed on these scenes with several analyses post hoc, and its brain gain was not measurable beyond low-level image features and replicated within only two of five withheld scenes (Sections J.1 and J.2).

Outlook. Designed routes, more participants and estimated camera paths can test single-trial decoding in new scenes, and generative 3D scene models could turn the DRAFT layout into complete room reconstructions.

AI USE STATEMENT

We used large language model assistants extensively, under the authors' supervision. They helped write and debug the analysis, model and evaluation code, develop the method, and propose, refine, preregister, run and audit experiments and analyses. They also helped interpret results, search the literature and verify references, translate the authors' notes into English, and draft and edit the text, equations and figures. They were not used to collect, generate or modify the fMRI data or the ground-truth geometry, no synthetic data enter any reported result, and the paper makes no theorem-style mathematical claims. The authors chose the research questions, checked all analyses, numbers, references and claims, and take full responsibility for this paper, including text, results and artifacts made with these tools.

ETHICS STATEMENT

This work analyzes existing fMRI recordings. The original study was approved by the institutional review board of the authors' institution, and all participants gave written informed consent. The name of the board and the protocol number are withheld for anonymous review. Analyses use coded participant identifiers and concern brain responses to rendered indoor scenes. The method depends on participant cooperation, known camera paths and repeated recordings, and it has not been tested for unconstrained or covert decoding. Neural data and derived readouts are sensitive.

REPRODUCIBILITY STATEMENT

Section 3 and Section A.1 describe the data, the cross-validation protocol and the metrics. Section 4 and Section A.2 give every model component with its hyperparameters, and Section B describes the statistical procedure. Every fitted parameter and every hyperparameter chosen by inner validation is selected without the held-out scene, and each model uses one fixed seed. Fixed design settings were chosen during development, which used all ten scenes. The clearest case is the range-calibration factor $a = 2$ (Section 4.3), fixed after earlier results on the same scenes (Section G). Nested selection by Chamfer distance picks the same value in all ten folds, and selection by F-score picks 2 or 2.5 (Section G). Decoder predictions were frozen before scoring under a written plan (Section C), and the transfer of the layout to five of the seven decoders is a post hoc analysis (Section C.2). One generator fold trains in about 15 minutes on one GPU (Section A.2). The supplementary material contains an anonymized static demo website that runs offline, with its source, and a short narrated video with subtitles that demonstrates it. The website holds an interactive viewer of the ten atlas routes, one per scene, chosen by a fixed rule, which shows predicted depth maps and merged 3D surfaces next to the ground truth. Upon publication, we will release the code for the method, trained models, evaluation manifests and derived readouts. Voxel-level fMRI data remain subject to the original consent and institutional data-sharing rules, and the derived readouts alone are not sufficient to retrain the cortical decoder.

REFERENCES

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

Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth Pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations (ICLR)*, 2025. URL `https://openreview.net/forum?id=aueXfY0Clv`.

Michael F. Bonner and Russell A. Epstein. Coding of navigational affordances in the human visual system. *Proceedings of the National Academy of Sciences*, 114(18):4793–4798, 2017. doi: 10.1073/pnas.1618228114.

Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 22710–22720, 2023a. doi: 10.1109/cvpr52729.2023.02175.

Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 24841–24858, 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4e5e0daf4b05d8bfc6377f33fd53a8f4-Abstract-Conference.html.

Rewon Child. Very deep VAEs generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=RLRXCV6DbEJ.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: Conditioning method for denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14347–14356, 2021. doi: 10.1109/iccv48922.2021.01410.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pp. 424–432. Springer International Publishing, 2016. doi: 10.1007/978-3-319-46723-8_49.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-scale embodied AI using procedural generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 5982–5994, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/27c546ab1e4f1d7d638e6a8dfbad9a07-Abstract-Conference.html.

Daniel D. Dilks, Joshua B. Julian, Alexander M. Paunov, and Nancy Kanwisher. The Occipital Place Area is causally and selectively involved in scene perception. *The Journal of Neuroscience*, 33 (4):1331–1336, 2013. doi: 10.1523/jneurosci.4081-12.2013.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/91c56ce4a249fae5419b90cba831e303-Abstract.html.

Alberto Elfes. Using occupancy grids for mobile robot perception and navigation. *Computer*, 22(6): 46–57, 1989. doi: 10.1109/2.30720.

Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998. doi: 10.1038/33402.

Russell A. Epstein, Eva Zita Patai, Joshua B. Julian, and Hugo J. Spiers. The cognitive map in humans: spatial navigation and beyond. *Nature Neuroscience*, 20(11):1504–1513, 2017. doi: 10.1038/nn.4656.

Oscar Esteban, Christopher J. Markiewicz, Ross W. Blair, Craig A. Moodie, A. Ilkay Isik, Asier Erramuzpe, James D. Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, Hiroyuki Oya, Satrajit S. Ghosh, Jessey Wright, Joke Durnez, Russell A. Poldrack, and Krzysztof J. Gorgolewski. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 16 (1):111–116, 2019. doi: 10.1038/s41592-018-0235-4.

Jianxiong Gao, Yuqian Fu, Yun Wang, Xuelin Qian, Jianfeng Feng, and Yanwei Fu. MinD-3D: Reconstruct high-quality 3D objects in human brain. In *Computer Vision – ECCV 2024*, pp. 312–329. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72970-6_18.

Guy Gaziv and Michal Irani. More than meets the eye: Self-supervised depth reconstruction from brain activity. *arXiv preprint arXiv:2106.05113*, 2021. doi: 10.48550/arXiv.2106.05113. URL https://arxiv.org/abs/2106.05113.

Zhanqiang Guo, Jiamin Wu, Yonghao Song, Jiahui Bu, Weijian Mai, Qihao Zheng, Wanli Ouyang, and Chunfeng Song. Neuro-3D: Towards 3D visual decoding from EEG signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23870–23880, 2025. doi: 10.1109/cvpr52734.2025.02223.

Linda Henriksson, Marieke Mur, and Nikolaus Kriegeskorte. Rapid invariant encoding of scene layout in human OPA. *Neuron*, 103(1):161–171.e3, 2019. doi: 10.1016/j.neuron.2019.04.014.

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

Joshua B. Julian, Jack Ryan, Roy H. Hamilton, and Russell A. Epstein. The Occipital Place Area is causally involved in representing environmental boundaries during navigation. *Current Biology*, 26(8):1104–1109, 2016. doi: 10.1016/j.cub.2016.02.066.

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

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9492–9502, 2024. doi: 10.1109/cvpr52733.2024.00907.

Mark D. Lescroart and Jack L. Gallant. Human scene-selective areas represent 3D configurations of surfaces. *Neuron*, 101(1):178–192.e7, 2019. doi: 10.1016/j.neuron.2018.11.004.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900, 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

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.

Shinji Nishimoto, An T. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646, 2011. doi: 10.1016/j.cub.2011.08.031.

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

Soojin Park, Timothy F. Brady, Michelle R. Greene, and Aude Oliva. Disentangling scene content from spatial boundary: Complementary roles for the Parahippocampal Place Area and Lateral Occipital Complex in representing real-world scenes. *The Journal of Neuroscience*, 31(4):1333–1340, 2011. doi: 10.1523/jneurosci.3885-10.2011.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671.

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

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12159–12168, 2021. doi: 10.1109/iccv48922.2021.01196.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2022. doi: 10.1109/tpami.2020.3019967.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, and Tanishq Abraham. Reconstructing the mind's eye: fMRI-to-Image with contrastive learning and diffusion priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 24705–24728, 2023. 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 *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44038–44059, 2024. URL https://proceedings.mlr.press/v235/scotti24a.html.

Anna Shafer-Skelton, Timothy F. Brady, and John T. Serences. A 2D Gabor-wavelet baseline model out-performs a 3D surface model in scene-responsive cortex. *PLOS Computational Biology*, 22 (2):e1013888, 2026. doi: 10.1371/journal.pcbi.1013888.

Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLOS Computational Biology*, 15(1):e1006633, 2019. doi: 10.1371/journal.pcbi.1006633.

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 (CVPR)*, pp. 14453–14463, 2023a. doi: 10.1109/cvpr52729.2023.01389.

Yu Takagi and Shinji Nishimoto. Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs. *arXiv preprint arXiv:2306.11536*, 2023b. URL https://arxiv.org/abs/2306.11536.

Erik Wijmans, Manolis Savva, Irfan Essa, Stefan Lee, Ari S. Morcos, and Dhruv Batra. Emergence of maps in the memories of blind navigation agents. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=lTt4KjHSsyl.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/s0893-6080(05)80023-1.

Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. DREAM: Visual decoding from reversing human visual system. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8211–8220, 2024. doi: 10.1109/wacv57701.2024.00804.

Xingqian Xu, Zhangyang Wang, Eric 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 (ICCV)*, pp. 7720–7731, 2023. doi: 10.1109/iccv51070.2023.00713.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10371–10381, 2024a. doi: 10.1109/cvpr52733.2024.00987.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 21875–21911, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/26cfdcd8fe6fd75cc53e92963a656c58-Abstract-Conference.html.

Taiping Zeng and Ming Bo Cai. Geometry representations along visual pathways in human spatial navigation. bioRxiv preprint, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824, 2023. doi: 10.1109/iccv51070.2023.00355.

Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6612–6619, 2017. doi: 10.1109/cvpr.2017.700.

APPENDIX CONTENTS

A Data and model details **16**

A.1 Data 16

A.2 Model components 16

A.3 When pooling across styles applies 18

B Evaluation and statistics **19**

B.1 Metric definitions 19

C Brain-decoding baselines **19**

C.1 DREAM depth branch 20

C.2 Post hoc layout transfer to all seven decoders 21

D Brain-input controls and cortical regions **22**

D.1 Camera motion, free space and EVC 23

D.2 Matched-budget comparison of cortical regions 23

D.3 Replacing the EVC features 24

D.4 Circular shift of the primary BOLD signal 24

D.5 Historical controls: single recordings and matched cohorts 25

E Per-scene results and atlas routes **25**

F Surface analyses **26**

F.1 Depth scale 26

F.2 Agreement across views 28

F.3 Pose drift during registration 30

G fMRI-free references, generator detail and range calibration **30**

H Loss correction and penalty ablation **33**

H.1 Penalty ablation 34

I Post hoc robustness analyses **35**

J Post hoc analyses: withheld scenes and low-level image features **37**

J.1 Scenes withheld from development 37

J.2 Low-level image features 40

K Additional analyses **41**

A DATA AND MODEL DETAILS

This section gives what is needed to reproduce the data preparation and every component of Sections 3 and 4. Section B defines the statistics and Section C the decoders. The next three sections support the main results: what the brain adds to the camera path (Section D), the layout gain in every scene (Section E) and the surfaces (Section F). Section G gives the fMRI-free references and the range calibration, and Section H the loss correction. Section I holds post hoc robustness analyses, and Section J the withheld scenes and low-level image features. Section K lists additional analyses of the same data that are not included in this version, each with its outcome. Throughout, DRAFT includes the range calibration of Section 4.3, and columns or rows marked $a = 1$ are its uncalibrated ablation (Section G).

A.1 DATA

Stimuli. Each video is an 80-second first-person walkthrough of a furnished indoor scene at 30 frames per second. The four styles of a scene (modern, Scandinavian, wabi-sabi and minimalist) share the floor plan and visit the rooms in the same order. Their camera paths are not frame-synchronized, so pooling across styles only pairs videos whose poses agree at the same second (Sections A.3 and 4.1). Seven scenes have all four styles, and the others two or three (Table 4).

Ground truth. For every second we cast rays from the recorded camera pose through a 384×512 pinhole grid with a 90° horizontal field of view and record the radial depth to the first surface of the 3D model, not the optical-axis depth. Depth is stored at 96×128 and averaged in inverse depth to 48×64. Depth Anything V2 predicts optical-axis depth z, converted to radial depth $r = z/u_z$ ($\boldsymbol{u}$: unit camera ray) before downsampling. Pixels closer than 5 cm or without hits are invalid.

fMRI. Blood-oxygen-level-dependent (BOLD) responses are resampled to the 2 mm MNI152NLin6Asym template and z-scored within each run. The readout uses V1 (3,004 voxels), V2 (2,020), ventral V3 (V3v, 866) and human V4 (hV4, 1,160). These are group regions from the Juelich maximum-probability atlas at a 25% threshold, not retinotopic maps measured in each participant. Targets are seconds 0 to 73, which leaves six seconds of video after the last target for the longest readout lag. Evaluation uses seconds 0 to 71, in nine non-overlapping eight-second routes. Run time series are linearly interpolated at stimulus-locked one-second centers, matching the repetition time (TR) of 1 s.

Cleaning. Before the stimulus windows are cut out, each run is detrended, z-scored and cleaned by regressing out 24 head-motion terms, cerebrospinal fluid and white-matter signals, five anatomical component-based noise correction (aCompCor) components and the available cosine terms. These head-motion terms differ from the camera-motion regressors of the readout. A run can contain several scenes, so this cleaning is not refit within each fold and can couple training and test responses. No depth target enters it. The results therefore describe offline decoding from run-cleaned responses, not independent preprocessing of a new scene.

Acquisition and participants. Data were acquired on a Siemens MAGNETOM Prisma at 3 T with a HeadNeck_20 coil (TR 1 s, echo time 30 ms, flip angle 55°, multiband factor 4, 60 slices of 2.5 mm) and preprocessed with fMRIPrep 24.1.1 (Nipype 1.8.6, ANTs 2.5.3, FreeSurfer 7.3.2) (Esteban et al., 2019). Scenes 5 to 7 each have 74 of 76, and scene 8 has 38 of 39 possible recordings. The records do not say whether a missing recording was never acquired, interrupted or excluded. Participant counts differ across scenes (6 to 19), so recording averages have unequal noise levels (matched-cohort controls in Section D.5).

A.2 MODEL COMPONENTS

Layout code. The 22 entries are 12 wall distances, four floor extents, four opening ratios, and the median and interquartile range (IQR) of log depth. Wall distances are the median log radial depth within 10° of the horizon in 12 bins of 7.5° over $[-45°, 45°]$. Floor extents are the 95th percentile of log depth over pixels below $-15°$ elevation in four vertical bands of equal width. Opening ratios are the fraction of valid pixels in each quadrant that are more than twice as far as the median depth

Table 4: **The ten scenes differ in styles and participants.** A recording is one participant watching one style video, and each video gives nine routes of eight consecutive one-second views. The participant total counts distinct people. Counts are those of DRAFT, and the decoders use two more recordings (scenes 9 and 10, Section C).

Scene	1	2	3	4	5	6	7	8	9	10	Total
Archive ID suffix	01	02	03	04	05	06	07	09	10	12	
Styles	4	4	4	4	4	4	4	3	2	3	36
Recordings	24	24	24	24	74	74	74	38	24	36	416
Participants	6	6	6	6	19	19	19	13	12	12	20
Routes	36	36	36	36	36	36	36	27	18	27	324

of the view. Empty wall or floor bins fall back to the median, and empty openings to zero. Targets are z-scored with the mean and standard deviation of the nine training scenes of each fold.

Readout. The readout averages six ridge regressions,

$$\hat{\boldsymbol{y}}_t = \tfrac{1}{6}\textstyle\sum_{m=1}^{6}\big(\boldsymbol{W}_m^\top[\boldsymbol{P}_m\boldsymbol{x}_t^{(m)};\ \boldsymbol{k}_t] + \boldsymbol{b}_m\big). \tag{5}$$

Here m indexes the members and t the view. $\boldsymbol{x}_t^{(m)}$ holds the 1,000 or 3,000 most reliable of the 7,050 voxels of early visual cortex (EVC). Their responses are averaged over 2 to 5 or 2 to 6 seconds after the view. $\boldsymbol{P}_m$ projects them onto 16 principal components, and $\boldsymbol{k}_t$ holds the 12 camera-motion regressors, also used at test time. The six members use (voxels kept, ridge penalty, lag window) of (3000, 0.1, 2 to 5 s), (1000, 0.3, 2 to 5 s), (3000, 0.3, 2 to 5 s), (1000, 0.1, 2 to 6 s), (1000, 0.1, 2 to 5 s) and (1000, 0.3, 2 to 6 s). Voxels are ranked by split-half reliability on the training scenes in the 3 to 6 s window. The ridge penalty is $\gamma N\boldsymbol{I}$ on z-scored features, where N is the number of training samples. The readout with a 3×3 target (Table 1) uses the same six ridges to predict the median log depth of each 3×3 block. It is a separate fit, not a rearrangement of the layout code.

Posterior spread. The posterior spread of an entry is the standard deviation of the readout's errors on a scene it was not trained on. To measure it without the held-out scene, we fit each readout on eight scenes, leaving out the held-out scene and one inner validation scene, and keep its errors on the validation scene. If the readout explains no training variance for an entry, its posterior mean and spread are the training mean and standard deviation. With pooling across styles, spreads are estimated separately for each number of pooled videos, or overall below 2,000 samples.

Renderer. The renderer uses the eye height and pitch of the camera at each second. Wall depth along a ray is $\exp(w(\theta))/\max(\cos\alpha, 0.2)$, where $w(\theta)$ interpolates the wall code at azimuth θ and α is the elevation after pitch rotation. The code is measured near the horizon, so this only approximates a planar wall. Floor depth is $h/(-d_y)$ for a ray with downward component $d_y < 0$ and eye height h, capped by the floor extent of its band. Openings blend each quadrant toward twice the median depth with a smooth step of half-width 6 pixels. Depth is clipped to $[0.05, 30]$ m. The uncertainty map carries the posterior spreads s_j of the code entries (Section 4.1) through the renderer. Let g be the dominant wall bin (or capped floor bin), with $s_g = 0$ on the purely geometric floor. Let q be the final inverse depth, q_o the inverse depth after opening blending, b the clipped global spread multiplier, and z the log-depth deviation before correction divided by its IQR. Then

$$(S^0)^2 = q^2\left[(bs_g)^2 + (zs_{21})^2 + s_{20}^2\right] + \left(\frac{q\,s_o}{\max(q_o, 10^{-6})}\right)^2, \quad s_o = |q_{\text{far}} - q_{\text{base}}|\sum_{k=1}^{4} M_k s_{16+k-1}. \tag{6}$$

Indices of s are zero-based, as in the code. q_{far} and q_{base} are the inverse depths of the opening target and of the unblended surface, and M_k are smooth quadrant masks. It omits covariance between entries, bin switches, quantiles and clipping, so the map is a reliability cue, not a calibrated posterior.

Generator. The generator is a 3D U-Net with 5.9M parameters and 48, 96 and 144 channels at three resolutions. It downsamples space twice but never the eight views. It uses residual blocks with group normalization, and two attention blocks at the lowest resolution apply spatial and temporal

self-attention (four heads). Its input channels are the flow state $\boldsymbol{z}_\tau$, the coarse depth $\boldsymbol{D}^0$ and the uncertainty map $\boldsymbol{S}^0$. The flow target $\boldsymbol{z}_1$ is the residual between the true inverse depth and $\boldsymbol{D}^0$, divided by one root-mean-square scalar ρ. The scalar ρ is computed on the training scenes and frozen for test time. The flow timestep and a 116-number vector condition the network through feature-wise linear modulation (FiLM), after a multilayer perceptron (MLP) of width 256. The vector holds the 22 posterior means and 22 posterior spreads, averaged over the eight views, and 72 motion features. They describe each of the eight views by its local position change, the sine and cosine of the heading change, pitch, eye-height change, speed and turning speed, and are standardized with fixed means and scales from 15,060 routes of ProcTHOR (Deitke et al., 2022), a set of synthetic houses. This is the only external data. No weights pretrained on the Natural Scenes Dataset (NSD) or on external depth data are used. We train with AdamW (learning rate 2×10^{-4}, weight decay 10^{-3}), gradient clipping at 1.0, an exponential moving average of weights with decay 0.999, batch size 16 and bf16 precision for 20,000 steps. One fold takes about 15 minutes and 3.7 GB of memory on one NVIDIA RTX 5090. Training uses all 67 overlapping routes of each training video (2,144 to 2,278 routes per fold), and testing the nine non-overlapping routes of each video. We draw eight samples per recording with 48 Euler steps and average all samples of a video in inverse depth. The reciprocal of this average, in meters, is $\hat{\boldsymbol{D}}_t$.

Coordinate-consistent flow penalty. The penalty of Equation (2) compares the posterior with the code of the generated depth. The map $\mathcal{A}_i(q) = (\mathcal{R}^+_{\text{aff}}(q) - \boldsymbol{\eta}_i)/\boldsymbol{\sigma}_i$ standardizes the fixed affine inverse of the renderer (Section H). The inner-fold mean $\boldsymbol{\eta}_i$ and scale $\boldsymbol{\sigma}_i$ standardize the posterior mean of sampled recording i. Weights are clipped precisions $w_j = \max(s_j, 10^{-3})^{-2}$, normalized by the batch mean, and $\lambda = 0.1$. The penalty is an approximation, not a calibrated likelihood.

Trajectory head and fusion head. For view t with position $\boldsymbol{p}_t$ and heading ψ_t, and for each of the 12 direction bins b of the layout code, we compute the farthest distance the camera reaches,

$$L_{t,b} = \max\big\{\|\boldsymbol{p}_{t'} - \boldsymbol{p}_t\| \,:\, |t' - t| \le K,\ \angle(\boldsymbol{p}_{t'} - \boldsymbol{p}_t) - \psi_t \in b\big\}, \tag{7}$$

which is zero for an empty bin. A distance reached by the camera is not an upper bound on the distance to a wall. For each bin, the free-space features store $\log(1 + L_{t,b})$, the number of path points in the bin and the time to the nearest one. They add the forward distance traveled over the previous and next K seconds (the signed displacement along the current heading) and the signed turn over the next four seconds, for 39 values. With the 72 motion features, the trajectory head has 111 inputs. Inner leave-one-scene-out validation chooses K from $\{4, 8, 16, 37, 73\}$ s and the ridge penalty from $\{1, 10, 10^2, 10^3, 10^4\}$. It maximizes layout r, with row order breaking ties. It selected $K = 4$ s in every scene we inspected, a penalty of 10^3 for the trajectory head in every scene, and for the fusion head 10^3 in eight scenes and 10^4 in scenes 6 and 10. The fusion head takes the 22 standardized pooled posterior means and the nine outputs of the trajectory head, standardized with training-scene statistics. The low-pass filter of Equation (4) is a separable Gaussian with $\sigma = 4$ pixels and symmetric padding, applied to log depth. The held-out scene is excluded throughout, but inner validation of the fusion head is not fully clean. It reuses frozen out-of-fold readout features, so the feature of one training scene can come from a readout trained on the inner validation scene. Removing this overlap would need a third level of folds that excludes all three scenes, which we did not fit. The trajectory head on motion features alone (inputs M in Table 1) is a frozen ridge on the 72 motion features. All layouts of Table 1 share the generator, which is conditioned on fMRI, so the surfaces of the trajectory head are not free of brain input.

A.3 When pooling across styles applies

Pooling helps only if partners exist and see similar depth, so we measure both. A video of another style is a partner at a second only if the two poses agree within 5 cm in each position coordinate, 2° in heading and 0.5°/s in turning speed. Pitch is not constrained, and the largest pitch difference is 1.092°. When several videos qualify, the one with the most recordings is the partner, ties broken by video name, never by scores. A partner is used at 1,473 of 2,592 video-seconds (56.8%), from 0% to 92.7% by scene, and never in scene 9. The depths of the two partners differ by a median of 0.314 m (IQR 0.172 to 0.563 m). This mixes small pose mismatch with changes in furniture.

B EVALUATION AND STATISTICS

The ten held-out scenes are the units of the bootstrap and sign-flip tests. Response windows of 2 to 6 s make neighboring views share BOLD samples, so per-view scores are not independent. For a mean metric such as layout r or pixel MAE, we take the difference between two methods on every route and average it within each scene. We then resample the ten scene differences with replacement 20,000 times and report the ten-scene mean with the 2.5th and 97.5th percentiles (paired scene bootstrap). For rate metrics such as row order, each scene pools its eligible cell pairs before differencing. Main-table layout and pixel-MAE point estimates instead pool all routes or pairs (route-pooled values), so they can differ slightly from ten-scene means. All intervals are descriptive 95% intervals without correction for multiple comparisons. All ten scenes were used during development, and each model was trained with one seed. Pose noise was tested at registration (Section F.3) and, post hoc, in the camera-path inputs (Section I). Single recordings with the corrected generator were tested only post hoc (Section I). Filmed real scenes, single-participant models and unobserved surfaces were not tested. Each analysis uses one fixed bootstrap seed, released with the code. Intervals of the historical models, trained before the loss correction (Section H), were route-weighted and are not combined with the corrected contrasts.

Sign-flip tests. Percentile intervals from ten scenes can be too narrow. For four main layout-r contrasts we therefore also run an exact two-sided sign-flip test. The ten scene differences are multiplied by all 2^{10} sign patterns, and p is the fraction of patterns whose absolute mean is at least the observed one. With ten scenes, the smallest possible value is 0.002. The contrasts are the DRAFT layout minus the generator alone ($p = 0.002$), minus the trajectory head ($p = 0.002$) and minus the fixed template ($p = 0.004$), and Brain-Diffuser + DRAFT layout minus Brain-Diffuser + trajectory layout ($p = 0.002$). After Holm correction for these four tests, the largest value is 0.008. The same test gives $p = 0.002$ for the four brain-input controls of Section 5.2: feature replacement in layout r and in F-score at $a = 2$, the circular shift in F-score at $a = 2$, and EVC minus the camera-motion regressors alone in cosine similarity. Each favors the intact input in all ten scenes. All these contrasts were named after the results were known and were not preregistered.

B.1 METRIC DEFINITIONS

Row order is the fraction of cell pairs in the same row of the 3×3 layout that the prediction orders correctly. Only pairs whose true log depths differ by at least 0.2 (about 22% in depth) count. Comparing cells within a row separates left-to-right structure from the near-floor and far-wall contrast that all indoor views share. Left/right uses the same threshold. Layout r is averaged over views and then over routes, so it measures agreement within views, not a pooled correlation. The surface code voxel-merges each route separately, but four routes at a time share a grid origin one meter below their joint minimum on each coordinate. The origin depends on the batch, so reproductions must keep the saved route order and batch size. Let P be the voxel-merged predicted points and Q the valid ground-truth ray hits from the same eight poses, not voxel-merged. With $d(p, Q) = \min_{q\in Q} \|p - q\|_2$,

$$C(P, Q) = \tfrac{1}{2}\left[|P|^{-1}\sum_{p\in P} d(p, Q) + |Q|^{-1}\sum_{q\in Q} d(q, P)\right], \tag{8}$$

$$p_\delta = |P|^{-1}\sum_{p\in P} \mathbf{1}[d(p, Q) \le \delta], \qquad r_\delta = |Q|^{-1}\sum_{q\in Q} \mathbf{1}[d(q, P) \le \delta], \tag{9}$$

$$F_\delta = \frac{2p_\delta r_\delta}{p_\delta + r_\delta}, \qquad \delta = 0.25\,\mathrm{m}. \tag{10}$$

The Chamfer distance C uses unsquared distances in meters. p_δ and r_δ are precision and recall, and F_δ is the F-score. Pixel MAE compares predicted radial depth with ground truth on valid pixels.

C BRAIN-DECODING BASELINES

All decoders share the folds, routes, targets and metrics of DRAFT, and every trainable part is fit on the nine training scenes of each fold.

Protocol. Decoders read participant-averaged responses, averaged over 2 to 6 seconds after each view, from seven regions of interest (ROIs). These are V1, V2, V3v and hV4 in early visual cortex and the scene-selective occipital place area (OPA), parahippocampal place area (PPA) and retrosplenial complex (RSC). A method that prescribes early visual areas uses V1 to hV4, and one that prescribes higher visual areas uses hV4, PPA, OPA and RSC. Each view's video frame is resized to the input size of each model without cropping. Reconstructed images are converted to metric depth with Depth Anything V2 (DA V2) Metric Indoor Large (Yang et al., 2024b). It overestimates depth on our rendered scenes by about a factor of 1.75, so we correct its scale with a log-affine map fitted on the frames of the training scenes (DA V2 scale calibration). Predictions were frozen before scoring under a written plan. The decoders read responses averaged over 418 recordings, and DRAFT over 416. The two extra recordings (one participant, minimalist style of scenes 9 and 10) have no second style to pool with. The trajectory layout uses no fMRI, and decoders with the DRAFT layout combine the two sets of responses.

Methods. *Linear depth decoding* maps fMRI to the first 64 principal components of log depth with ridge regression (Hoerl & Kennard, 1970), a standard linear approach to fMRI decoding (Naselaris et al., 2011). *CLIP retrieval* maps fMRI to the image embedding of CLIP (Contrastive Language-Image Pre-training) ViT-L/14 (Radford et al., 2021) with ridge regression, a simple analogue of the contrastive retrieval branch of MindEye (Scotti et al., 2023), and averages the depth maps of the ten nearest training frames. *Image decoding* follows Takagi & Nishimoto (2023a). Early visual cortex predicts the Stable Diffusion latent, and higher visual cortex predicts the text embedding of a caption. Takagi & Nishimoto (2023a) averaged the five human-written captions of each image in the COCO dataset. Video frames have no captions, so we instead caption each frame with BLIP, an image captioning model (Li et al., 2022). Image-to-image diffusion then generates the image (strength 0.8, 50 denoising diffusion implicit model steps, guidance 5.0). *Depth decoding* (Takagi & Nishimoto, 2023b) predicts hidden states of the dense prediction transformer DPT-large (Ranftl et al., 2021), whose head outputs relative depth calibrated on the training scenes. *Brain-Diffuser* (Ozcelik & VanRullen, 2023) predicts latents of a very deep variational autoencoder (VDVAE) (Child, 2021) and CLIP image and text embeddings that guide Versatile Diffusion (Xu et al., 2023) (strength 0.75, mixing 0.4). It keeps its released step that re-scales its predictions to their mean and spread over the held-out videos. This step uses no labels, but it uses all held-out predictions at once (it is transductive). *MindEye* (Scotti et al., 2023) trains an MLP (width 1024, 64 epochs per fold) to predict CLIP ViT-L/14 hidden states, and a low-level branch starts Versatile Diffusion image variation (strength 0.85). Our version omits the ConvNeXt contrastive term and uses the Versatile Diffusion autoencoder for low-level targets. We did not reproduce Gaziv & Irani (2021), full DREAM (Xia et al., 2024) or MinD-Video. Our adaptations decode each view alone, so they cannot show that DRAFT beats temporal fMRI-to-video decoders.

Regularization. On the penalty grids of Takagi & Nishimoto (2023a) and of depth decoding, every fold chose the largest penalty, so the best one could lie outside these grids. With these penalties image decoding also produced textureless color fields. We therefore report the stronger "wider ridge" variants, fit on a wider grid. For linear depth, a wider grid selected the same penalty.

Image similarity. We test whether CLIP ViT-L/14 embeddings pick the viewed frame over another frame of the same held-out scene. This two-way identification is 0.500 and 0.525 for the official and wider-ridge versions of Takagi & Nishimoto (2023a), 0.627 for Brain-Diffuser and 0.624 for MindEye (chance 0.5). The grayscale structural similarity (SSIM) of the wider-ridge decoded images to the viewed frame is 0.651, against 0.649 for another frame of the same scene. So the images are indoor scenes that match their specific view weakly at best (Section K).

C.1 DREAM DEPTH BRANCH

We adapt the released depth-regression branch of DREAM (Xia et al., 2024), which uses a VDVAE pretrained on ImageNet. In each fold, ridge regression (penalty fixed at 50,000) maps the fMRI of the training scenes to VDVAE latents of depth images, coded as gray level $\log(d/0.05)/\log(600)$ at 64×64. Eight decoded samples are averaged per view. Two differences remain. The DREAM depth branch has no camera input, and it reads the decoders' responses, which average two more recordings. The primary contrast, fixed in advance, is the DREAM depth branch minus the generator

alone in Chamfer distance at $a = 1$. It is $+0.134$ m [0.097, 0.175], and all ten scenes favor the generator. F-score is also lower in all ten scenes. The DREAM depth branch has the better layout r ($+0.128$ [0.087, 0.158], nine scenes) and is better on most other depth-map layout measures (Section K). The differences in input and cohort prevent attributing the gap to the architecture.

Range ratio. The range ratio measures how much of the true depth spread a prediction keeps (Section 5.4). It divides the standard deviation of the predicted 3×4 cell log depths by that of the true ones, so its ideal value is 1. Before range calibration, the route-pooled range ratios of our ridge layouts are 0.46 for the readout with a 3×3 target, 0.50 for the trajectory head from motion, 0.54 for the trajectory head with free space and 0.49 for DRAFT. The ratio is 0.78 for the generator, and DA V2 gives Brain-Diffuser and MindEye 0.84. Range calibration raises the ratio of DRAFT to 0.97. Section K summarizes ten-scene means instead.

C.2 Post hoc layout transfer to all seven decoders

After the two planned combinations (Section G) were scored, we asked whether the layout guides every decoder and applied the same low-frequency replacement (Equation (4)) to the frozen depth of all seven decoders, with the trajectory layout or the DRAFT layout, and scored each version at $a = 1$ and $a = 2$ (every score and paired difference omitted; Section K). The transfer's Brain-Diffuser and linear-depth rows reproduce the planned combinations exactly, and its $a = 2$ versions form the "+ path" and "+ DRAFT" columns of Table 2.

With the DRAFT layout, every decoder reaches a layout r of 0.487 to 0.493, whatever its own value (0.171 to 0.429). Over the trajectory layout, the DRAFT layout raises layout r by 0.036 to 0.041 and lowers pixel MAE by 0.077 to 0.078 m, in all ten scenes for each decoder (sign-flip $p = 0.002$ each). Row order and left/right gain less (0.007 to 0.015), and no sign-flip p is below 0.05.

The surfaces are mixed. At $a = 1$, the replacement lowers the F-score of the three decoders whose surfaces are best alone (Brain-Diffuser, MindEye and CLIP retrieval) in at least eight of ten scenes each, and raises their Chamfer distance in at least seven. It helps the two with the weakest surfaces, depth decoding and the DREAM depth branch. At $a = 2$, every decoder except linear depth has a higher F-score with the DRAFT layout than alone. But the factor was fixed for the layout rows, and it hurts the unmodified decoders. Unmodified Brain-Diffuser, for example, has an F-score of 0.492 at $a = 1$ and 0.424 at $a = 2$. So these contrasts do not show that the layout beats each decoder at its own best factor. For Brain-Diffuser, that comparison is within noise (Section G). The brain part of the DRAFT layout adds little to the surfaces. Over the trajectory layout, the F-score changes by $+0.017$ to $+0.020$ at $a = 1$, with every interval including zero, and by $+0.014$ to $+0.023$ at $a = 2$. The Chamfer distance at $a = 2$ falls by 0.018 to 0.030 m in seven to nine scenes, but the sign-flip p is above 0.05 for linear depth, depth decoding and the DREAM depth branch. The exact sign-flip test is the conservative reading.

Panel (a) of Table 5 compares DRAFT itself with each decoder under the three conditions of Table 2. Given the same camera path, that is, with the trajectory layout, every decoder has a lower layout r than DRAFT, in eight to ten scenes (sign-flip $p \leq 0.010$). Linear depth, image decoding, depth decoding and the DREAM depth branch also have a lower F-score (eight to ten scenes, $p \leq 0.016$). The F-scores of CLIP retrieval, Brain-Diffuser and MindEye are slightly lower but within noise (six scenes each, $p \geq 0.22$). We added these sign-flip tests after the results were known.

Panel (b) of Table 5 gives the gains behind the markers of Table 2. Over its unmodified output, every decoder has a positive mean gain with the DRAFT layout in all three metrics. The intervals exclude zero in 17 of 21 cases (sign-flip $p \leq 0.037$). The four exceptions are the F-score and Chamfer distance of Brain-Diffuser and MindEye, whose gains are small ($p \geq 0.35$). These gains include range calibration, which the unmodified output lacks. When both are calibrated, the DRAFT layout does not raise the F-score of linear depth over the decoder alone (-0.002 at $a = 2$, Section K). Table 2 omits pixel MAE. Table 1 gives it for DRAFT. It is 0.501 m for depth from the video frames and 1.188 m for the fixed template.

Table 5: **Given the same camera path, DRAFT has a higher layout r than every decoder, and the DRAFT layout improves every decoder.** (a) DRAFT minus each method of Table 2. (b) The markers of Table 2: decoder with the DRAFT layout and range calibration ($a = 2$) minus unmodified decoder ($a = 1$), * if the interval excludes zero. The surface gains of Brain-Diffuser and MindEye are not reliable. Ten-scene means of per-scene differences, 95% paired scene-bootstrap intervals and scenes of ten favoring DRAFT or the layout. Post hoc exact sign-flip p (Section B) for (b) and the trajectory-layout rows of (a). Layout rows are post hoc except for linear depth and Brain-Diffuser (Section C.2). ‡DA V2 on the frames. †Our adaptation (Section C.1).

(a) DRAFT minus each method

	Layout $r\uparrow$			F@25cm↑		
DRAFT minus	Mean [95% CI] (scenes)		p	Mean [95% CI] (scenes)		p
References						
Video frames + DA V2‡	−0.393 [−0.445, −0.340]		–	−0.258 [−0.295, −0.221]		–
Fixed template	+0.144 [0.084, 0.204]		–	+0.093 [0.080, 0.108]		–
Decoder alone (unmodified output)						
Linear depth	+0.045 [0.011, 0.079]	(7)	–	+0.091 [0.066, 0.113]	(10)	–
CLIP retrieval	+0.210 [0.157, 0.267]	(10)	–	+0.029 [0.014, 0.044]	(8)	–
Image decoding	+0.300 [0.259, 0.335]	(10)	–	+0.042 [0.021, 0.062]	(9)	–
Depth decoding	+0.156 [0.123, 0.193]	(10)	–	+0.167 [0.136, 0.194]	(10)	–
DREAM depth branch†	+0.045 [0.012, 0.075]	(7)	–	+0.116 [0.086, 0.143]	(10)	–
Brain-Diffuser	+0.152 [0.128, 0.175]	(10)	–	−0.008 [−0.027, 0.011]	(4)	–
MindEye	+0.206 [0.168, 0.249]	(10)	–	−0.010 [−0.023, 0.004]	(4)	–
+ trajectory layout ("+ path"), the same camera path as DRAFT						
Linear depth	+0.028 [0.016, 0.039]	(9)	0.006	+0.044 [0.028, 0.060]	(10)	0.002
CLIP retrieval	+0.031 [0.017, 0.044]	(8)	0.008	+0.011 [−0.005, 0.026]	(6)	0.223
Image decoding	+0.032 [0.020, 0.044]	(10)	0.002	+0.028 [0.012, 0.043]	(8)	0.016
Depth decoding	+0.027 [0.015, 0.038]	(9)	0.006	+0.049 [0.032, 0.065]	(10)	0.002
DREAM depth branch†	+0.032 [0.020, 0.044]	(9)	0.004	+0.049 [0.032, 0.065]	(10)	0.002
Brain-Diffuser	+0.024 [0.011, 0.037]	(8)	0.010	+0.008 [−0.007, 0.021]	(6)	0.348
MindEye	+0.031 [0.018, 0.045]	(9)	0.006	+0.008 [−0.006, 0.023]	(6)	0.328
+ DRAFT layout						
Linear depth	−0.009 [−0.013, −0.006]	(1)	–	+0.029 [0.024, 0.033]	(10)	–
CLIP retrieval	−0.008 [−0.014, −0.002]	(2)	–	−0.009 [−0.014, −0.004]	(2)	–
Image decoding	−0.008 [−0.013, −0.003]	(1)	–	+0.010 [0.006, 0.014]	(9)	–
Depth decoding	−0.011 [−0.016, −0.006]	(1)	–	+0.035 [0.029, 0.041]	(10)	–
DREAM depth branch†	−0.008 [−0.012, −0.004]	(1)	–	+0.035 [0.030, 0.040]	(10)	–
Brain-Diffuser	−0.012 [−0.018, −0.007]	(1)	–	−0.015 [−0.022, −0.007]	(1)	–
MindEye	−0.007 [−0.013, −0.001]	(2)	–	−0.015 [−0.022, −0.008]	(1)	–

(b) Decoder + DRAFT layout ($a = 2$) minus the decoder alone ($a = 1$), the markers of Table 2

	Layout $r\uparrow$			F@25cm↑			Chamfer (m)↓		
Decoder	Gain [95% CI] (scenes)		p	Gain [95% CI] (scenes)		p	Gain [95% CI] (scenes)		p
Linear depth	+0.054* [0.021, 0.088]	(7)	0.020	+0.062* [0.041, 0.081]	(9)	0.004	+0.072* [0.029, 0.109]	(9)	0.010
CLIP retrieval	+0.219* [0.166, 0.274]	(10)	0.002	+0.038* [0.023, 0.053]	(10)	0.002	+0.038* [0.012, 0.059]	(9)	0.020
Image decoding	+0.308* [0.268, 0.343]	(10)	0.002	+0.031* [0.009, 0.053]	(7)	0.037	+0.042* [0.017, 0.067]	(7)	0.016
Depth decoding	+0.166* [0.134, 0.204]	(10)	0.002	+0.132* [0.104, 0.155]	(10)	0.002	+0.283* [0.219, 0.344]	(10)	0.002
DREAM depth branch†	+0.053* [0.021, 0.083]	(7)	0.025	+0.081* [0.054, 0.105]	(9)	0.004	+0.115* [0.065, 0.162]	(9)	0.006
Brain-Diffuser	+0.164* [0.140, 0.188]	(10)	0.002	+0.006 [−0.009, 0.022]	(5)	0.457	+0.011 [−0.012, 0.032]	(7)	0.354
MindEye	+0.213* [0.171, 0.259]	(10)	0.002	+0.005 [−0.007, 0.017]	(5)	0.486	+0.005 [−0.020, 0.025]	(8)	0.713

D BRAIN-INPUT CONTROLS AND CORTICAL REGIONS

These controls ask what the brain adds to camera motion, which cortical input carries layout information and whether the pipeline depends on it. None tests biological causation.

D.1 CAMERA MOTION, FREE SPACE AND EVC

One test of what EVC adds to camera motion keeps the motion input fixed. We fit the readout with a 3×3 target, on the same folds, to the 12 camera-motion regressors alone (motion alone). It reaches a layout r of 0.338 (route-pooled), and adding EVC raises it by 0.098 [0.053, 0.151], in nine scenes ($p = 0.004$). A second comparison starts from the trajectory head on the 72 motion features alone (layout r 0.390, Table 1). Two alternatives beat it by the same amount, each in eight scenes: the readout with a 3×3 target from EVC and the 12 regressors, by 0.050 [0.015, 0.092] ($p = 0.025$), and the trajectory head with free space added to the motion features, by 0.050 [0.022, 0.076] ($p = 0.012$). The readout also lowers pixel MAE by 0.117 m [0.061, 0.172] in eight scenes ($p = 0.010$). The two read different motion inputs, so this comparison alone does not isolate EVC.

D.2 MATCHED-BUDGET COMPARISON OF CORTICAL REGIONS

To test the choice of EVC (Section 4.1), we compare EVC (V1, V2, V3v and hV4), scene-selective regions (OPA, PPA and RSC) and their union under one readout fixed before scoring: 1,000 most reliable voxels, 16 principal components, ridge penalty $\gamma = 0.1$ and a 2 to 5 s lag window. Voxel ranking, principal components and standardization use only the nine training scenes of each fold. Every readout includes the 12 camera-motion regressors, and a ridge on these alone is the reference. The comparison tests the readout alone, without a noise ceiling.

EVC predicts the code best (Table 6). In cosine similarity, it beats the scene-selective regions by 0.160, in nine scenes, and these still beat motion alone by 0.059, also in nine scenes (intervals in Tables 3 and 6). Adding the scene-selective regions to EVC changes little: the union minus EVC is -0.0003 [-0.0030, 0.0018] in cosine similarity, and the union's mean squared error is lower by only 0.0010 [0.0002, 0.0021]. This is expected, because the union's reliability ranking keeps 979 to 992 EVC voxels and only 8 to 21 scene-selective voxels per fold (PPA 5 to 18, OPA 0 to 3, RSC 0 to 2). The union therefore does not test a separate budget for the scene-selective regions, and no comparison shows that a region is necessary. For every input, wall distances are the best predicted of the four groups of entries, a property of these targets and this readout.

The responses used here average 25 and 37 recordings in scenes 9 and 10, against 24 and 36 for DRAFT, so this is not a full-pipeline ablation with one cohort. Camera-motion regressors do not control image motion from texture, occlusion and parallax, so it shows predictive information, not a 3D representation independent of image features (Shafer-Skelton et al., 2026).

Lag profile and circular-shift control. A gain carried by the BOLD response should peak a few seconds after the stimulus and vanish when the BOLD series is shifted. We refit the readout with single-second lags from 0 to 6 s and, for the 2 to 5 s window, after circular shifts of the BOLD series by 20 offsets from 10 to 70 s (Figure 5). The gain over motion alone is largest at 3 and 4 s, in all ten scenes (0.205 [0.143, 0.266] and 0.204 [0.135, 0.271]). It is smallest at 0 s, at 0.054 [0.010, 0.091], in eight scenes. Some gain at 0 s is expected, because the layout code changes slowly and the hemodynamic response is spread in time. The unshifted gain (0.219) exceeds all 20 shifted gains (largest 0.077), a permutation p of 1/21. The pooling partner does not enter this readout.

Table 6: **With 1,000 voxels each, EVC predicts the layout code best, and the scene-selective regions beat motion alone.** Ten-scene means. Cos.: cosine similarity of centered predicted and true codes. Entry r: mean per-entry correlation (neither is layout r). MSE, MAE: mean squared and absolute error (std.: standardized units). ΔCos.: minus motion alone, with 95% paired scene-bootstrap interval. Scenes > 0: of 10, not applicable to the reference row. Bold: within 0.002 of best. Contrasts between regions: Section D.2 and Table 3.

Input	Cos. ↑	Entry r ↑	MSE ↓ (std.)	MAE ↓ (std.)	ΔCos. ↑	Scenes > 0
EVC (V1 to hV4)	**0.334**	**0.394**	**0.848**	**0.678**	+0.219 [+0.150, +0.290]	10
Scene-selective	0.174	0.221	0.971	0.740	+0.059 [+0.019, +0.088]	9
Union (both)	**0.334**	**0.394**	**0.847**	**0.678**	+0.219 [+0.151, +0.287]	10
Motion alone (reference)	0.115	0.185	0.987	0.751	–	–

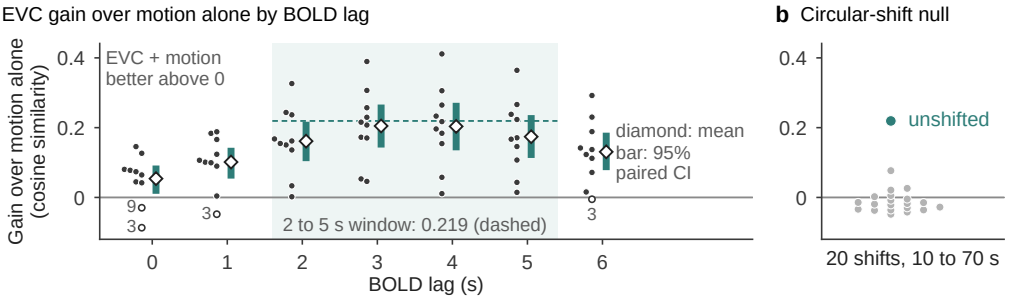


Figure 5: **EVC's gain over camera motion peaks at a 3 to 4 s lag and exceeds all 20 circular shifts.** Gain: cosine similarity of the EVC readout minus that of motion alone. **(a)** Single-second BOLD lags. Dots: scenes (hollow and numbered at or below zero). Diamonds and bars: mean and 95% paired scene-bootstrap interval. Shading and dashed line: the 2 to 5 s window of the main readout, which averages four lags and can gain more than any single one. **(b)** Circular shifts of 10 to 70 s (gray) and the unshifted window (teal), on the scale of (a).

Table 7: **Replacing or shifting the brain input worsens layout and surfaces in most scenes.** Intact minus control, for feature replacement (readout kept) and circular shift (readout refitted). Ten-scene means with 95% paired scene-bootstrap intervals (low, high). Scenes better: of ten, scenes in which the intact input scores better (per-scene values omitted; Section K). Pixel MAE and Chamfer in cm (m in the other tables), F in percentage points (pp). G: generator, D: DRAFT, subscript: a. Only one mean favors the control, G_2 Chamfer under feature replacement, within noise.

	Layout r ↑		Pixel MAE (cm) ↓				F@25cm (pp) ↑				Chamfer (cm) ↓			
	G	D	G_1	G_2	D_1	D_2	G_1	G_2	D_1	D_2	G_1	G_2	D_1	D_2
Feature replacement: intact minus centered EVC features														
Mean	+0.064	+0.037	−5.7	−3.5	−7.7	−4.2	+2.1	+1.1	+4.0	+3.6	−2.3	+0.3	−4.7	−2.5
95% low	+0.046	+0.025	−7.4	−5.8	−10.6	−6.6	+1.4	+0.7	+2.2	+2.4	−3.0	−0.4	−6.4	−3.6
95% high	+0.082	+0.051	−4.4	−1.4	−5.3	−2.0	+2.8	+1.5	+5.9	+5.0	−1.7	+0.9	−3.1	−1.1
Scenes better	10	10	10	7	10	9	10	10	10	10	10	4	10	9
Circular shift: intact minus shifted primary BOLD														
Mean	+0.078	+0.043	−6.9	−6.9	−8.4	−6.8	+2.5	+1.6	+4.2	+4.4	−2.8	−0.8	−5.2	−3.6
95% low	+0.047	+0.028	−9.1	−9.5	−11.8	−10.6	+1.6	+1.1	+2.5	+2.9	−3.8	−1.4	−7.0	−5.3
95% high	+0.112	+0.059	−5.0	−4.4	−5.6	−4.2	+3.3	+2.2	+6.1	+6.0	−2.0	−0.1	−3.5	−2.0
Scenes better	10	10	10	10	10	10	10	10	10	10	10	8	10	9

D.3 Replacing the EVC features

To test whether the fitted pipeline uses the EVC features it receives at test time, for each recording and its pooling partner we set the scores of the EVC principal components to their training-scene center and keep the fitted readout. Camera-motion regressors, model weights, pooling partners, posterior spreads, trajectory layouts and reliability masks stay fixed. We recompute the posterior means and everything that depends on them: coarse depth, uncertainty maps and DRAFT layouts. With intact (unaltered) features, the layout r of DRAFT is higher by 0.0373 and its F-score at $a = 2$ by 0.0363, in all ten scenes. Its pixel MAE falls by 0.0418 m and its Chamfer distance by 0.0251 m, both in nine scenes (intervals in Table 7). In scene 1, pixel MAE and Chamfer distance at $a = 2$ are worse, although layout r and F-score improve. For the generator alone, the Chamfer change at $a = 2$ is +0.0028 m [−0.0037, 0.0091], within noise. This control is not a shuffle of the raw BOLD signal. The posterior spreads keep information from the intact fit, and centered features can combine unusually with the unchanged camera-motion regressors.

D.4 Circular shift of the primary BOLD signal

Feature replacement keeps the fitted readout. To test whether the full pipeline also depends on the timing of the BOLD signal, we circularly shift the primary BOLD signal (that of the evaluated

Table 8: **DRAFT raises layout r in every held-out scene, and its surface comparison with the generator depends on a.** At $a = 2$ it lowers pixel MAE in every scene and Chamfer distance in eight, but its F-score is higher in only four. Scene means (equal weights, 324 routes) of the generator (G) and DRAFT (D). Pixel MAE and surfaces at $a = 2$. D−G: mean difference, 95% paired scene-bootstrap interval and scenes with D better. Bold: better mean. At $a = 1$ the F-score and Chamfer distance of DRAFT are worse than those of the generator in every scene (Table 13, Section K).

		Held-out scene													D
		1	2	3	4	5	6	7	8	9	10	Mean	D−G	95% interval	better
Layout r ↑	G	0.307	0.336	0.314	0.283	0.343	0.277	0.285	0.275	0.262	0.324	0.301			
	D	0.551	0.513	0.578	0.532	0.492	0.359	0.479	0.391	0.408	0.427	**0.473**	+0.172	[+0.135, +0.210]	10/10
Pixel MAE (m) ↓	G	1.419	1.493	1.577	1.434	1.529	1.559	1.410	1.512	1.668	1.572	1.517			
	D	1.225	1.131	1.049	1.024	1.242	1.153	1.120	1.172	1.289	1.211	**1.162**	−0.356	[−0.408, −0.304]	10/10
F@25cm ↑	G	0.455	0.478	0.465	0.480	0.534	0.506	0.472	0.462	0.466	0.464	0.478			
	D	0.447	0.507	0.530	0.534	0.496	0.504	0.481	0.456	0.441	0.443	**0.484**	+0.006	[−0.013, +0.026]	4/10
Chamfer (m) ↓	G	0.428	0.457	0.451	0.425	0.400	0.478	0.458	0.502	0.452	0.437	0.449			
	D	0.446	0.378	0.339	0.349	0.412	0.415	0.406	0.437	0.435	0.433	**0.405**	−0.044	[−0.069, −0.018]	8/10

recording) by 37 of 80 positions before lag averaging and fit a separate principal component analysis and ridge readout on the shifted data. The camera-motion regressors are not shifted. Voxel rankings, posterior spreads and masks from the intact data are kept, and we regenerate every dependent prediction on 324 matched eight-view routes. The primary contrast, fixed in advance, is the F-score of DRAFT at $a = 2$. The intact signal raises it from 0.4403 to 0.4840 (by 0.0436) in all ten scenes (interval in Table 3). Chamfer distance favors the intact signal on average but not in every scene. At $a = 2$, intact BOLD gives the generator a worse Chamfer distance in scenes 2 and 9, and DRAFT a worse one in scene 1. At $a = 1$, both improve in all scenes. Layout r does not depend on a, so its two contrasts are one result. Because the readout is refitted and information from the intact data remains, this is not a complete null control. Table 7 summarizes both controls.

D.5 Historical controls: single recordings and matched cohorts

Two earlier controls agree with the corrected ones: the intact signal still helps when fewer recordings are averaged at inference. Both used the historical generator (before the loss correction, Section H), no pooling across styles and the same 37-position shift. With one recording per reconstruction (readout still trained on averaged data), intact BOLD raises F-score by 0.0123 [0.0062, 0.0187] and lowers Chamfer distance by 0.0157 m [0.0076, 0.0236]. With matched cohorts (the three participants present in all ten scenes, and six participants matched within each of three acquisition blocks, with equal block weights as planned), it raises F-score by 0.0115 [0.0057, 0.0179] and lowers Chamfer distance by 0.0153 m [0.0077, 0.0226]. Neither control measures data efficiency, because the cohorts change the evaluated recordings, not the training budgets. Section I analyzes single recordings with the corrected generator (post hoc). Across scenes, the number of recordings (Table 4) is not clearly related to the feature-replacement effect on DRAFT (Spearman correlation 0.14 for layout r and -0.31 for the F-score at $a = 2$, ten scenes, descriptive only).

E Per-scene results and atlas routes

Table 8 compares the generator and DRAFT in each held-out scene. Its means weight scenes equally, so they differ slightly from the route-pooled values of Tables 1 and 13. Table 13 gives every method at $a = 1$ and $a = 2$.

Route selection. The atlas has one route per scene: the lower-median route by the uncalibrated layout r of DRAFT, with ties broken by route order. We froze this rule before rendering, so it does not depend on surface quality. Figures 4 and 6 instead show display routes: among the 144 where DRAFT and Brain-Diffuser + DRAFT layout both beat unmodified Brain-Diffuser in F-score, the highest by the F-score of DRAFT (Figure 4: clearest of the top five). In Figure 4, 39% of the ground-truth points lie above 2.4 m or outside the plotted area. As in the surface atlas (Section K), such points are hidden but scored (at most 3,907 predicted and 227 ground-truth points per panel). Unmodified Brain-Diffuser has slightly better mean surface scores at $a = 1$ than DRAFT at $a = 2$

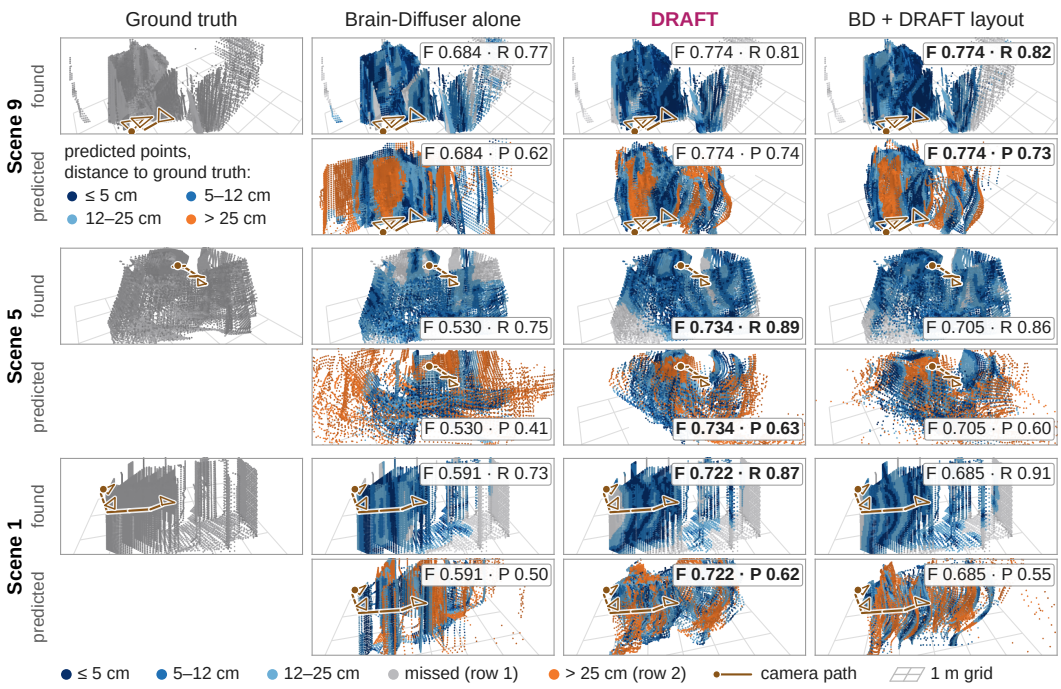


Figure 6: **Three routes chosen for display, not representative: predictions recover the room's extent and main wall positions, not clear walls or furniture.** Rule: among routes where DRAFT and Brain-Diffuser + DRAFT layout beat unmodified Brain-Diffuser (Brain-Diffuser alone) in F-score, highest DRAFT F-score first, one per scene, without the route of Figure 4 (rule-chosen routes: Table 9). Brain-Diffuser alone: Table 2. Row 1 (found): true points by distance to the nearest prediction (R: recall). Row 2 (predicted): predicted points by distance to the ground truth (P: precision). Points above 2.4 m or outside the plotted area, and in row 2 stray points (fewer than 4 neighbors or clusters under 40 points, radii scaled with range), are hidden for display, but all points are scored. F: route F-score (bold: best of three).

(Table 13). Figure 7 shows all eight depth views of the atlas routes of scenes 1 to 4 for DRAFT, unmodified Brain-Diffuser and Brain-Diffuser + DRAFT layout, with route layout r.

Failures. In the route of scene 5, fusion raises layout r from 0.419 to 0.461, yet the F-score falls from 0.659 to 0.328. In scene 4 it instead rises from 0.536 to 0.686. One route can mislead either way, so we report route views (Figure 7) and aggregate scores.

F Surface analyses

This section tests three explanations of the surface scores: depth range (Section F.1), agreement across views (Section F.2) and pose drift (Section F.3).

F.1 Depth scale

To test whether the surface gap between methods is only a matter of depth range, we align each predicted view to its ground truth in log depth before lifting. Median alignment matches the medians (Zhou et al., 2017). The log-affine alignment also fits a slope, so it undoes range calibration as well (compare Ranftl et al., 2022). A least-squares slope shrinks noisy predictions toward the mean. This fit therefore partly measures per-view correlation again, and even the fixed template reaches F-score 0.575 under it. Both alignments use test depth and are post hoc. We also report three unaligned pixel metrics: the scale-invariant log error (SI-log, $\times 100$) (Eigen et al., 2014), the mean absolute relative error (AbsRel), and the share of pixels within a factor of 1.25 of the true depth ($\delta < 1.25$).

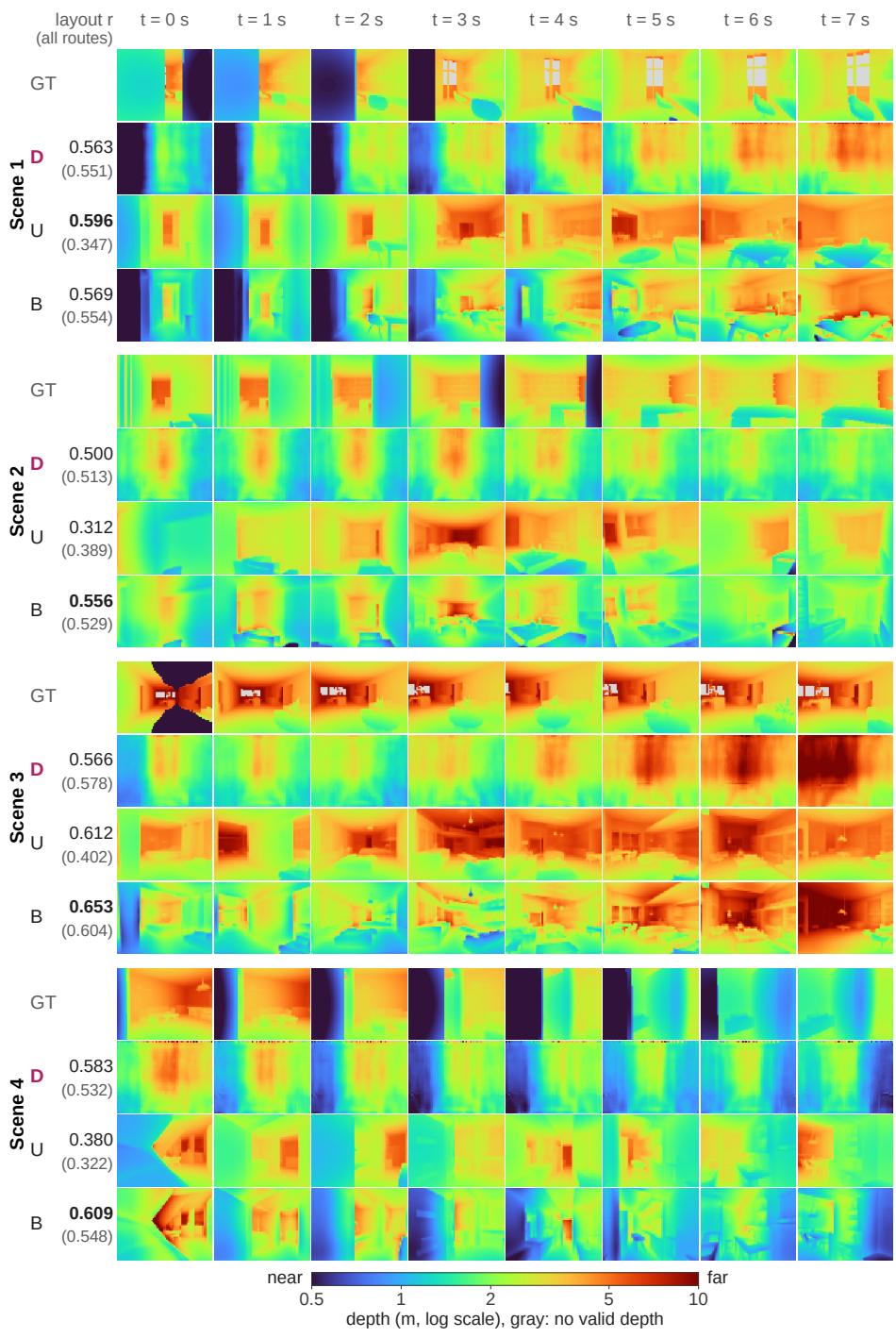


Figure 7: **Depth atlas, scenes 1 to 4, failures included: on the ten rule-chosen routes, Brain-Diffuser + DRAFT layout has the highest route layout r in six scenes, DRAFT in three and unmodified Brain-Diffuser in one.** All eight views, 1 s apart, of the rule-chosen routes of Table 9 (Figure 1 is hand-picked). Rows: furnished ground truth (GT, gray: invalid), DRAFT (D, $a = 2$), unmodified Brain-Diffuser (U, calibrated DA V2 depth, $a = 1$) and Brain-Diffuser + DRAFT layout (B, $a = 2$). Numbers: route layout r (bold: best of three) and the scene's value over all routes (in parentheses). Means over these ten and all 324 routes: 0.493 and 0.480 (D), 0.236 and 0.324 (U), 0.509 and 0.493 (B). Scenes 5 to 10: Section K.

Table 9: **On the ten atlas routes DRAFT improves the layout, but its F-score is the lowest of the three methods in seven routes.** Scores of the ten atlas routes, whose surface atlas is listed in Section K, all at $a = 2$. G: generator, D: DRAFT, BD + L: Brain-Diffuser + DRAFT layout (surfaces only). Bold per row: best of three for surfaces. Mean: over the ten routes. Bottom: paired differences with 95% route-bootstrap intervals and routes where D is better.

	Depth map				Merged eight-view surface					
	Layout $r\uparrow$		Pixel MAE (m)$\downarrow$		F@25cm$\uparrow$			Chamfer (m)$\downarrow$		
Scene	G	D	G	D	G	D	BD + L	G	D	BD + L
1	0.536	0.563	0.988	0.856	**0.574**	0.541	0.488	**0.277**	0.279	0.304
2	0.356	0.500	1.258	0.846	0.560	0.550	**0.606**	0.426	0.299	**0.252**
3	0.010	0.566	2.277	1.756	**0.435**	0.431	0.434	0.484	0.440	**0.403**
4	0.435	0.583	1.221	0.705	0.536	**0.686**	0.635	0.399	**0.238**	0.259
5	0.419	0.461	1.288	1.349	**0.659**	0.328	0.415	**0.291**	0.448	0.371
6	0.065	0.400	1.623	1.188	0.442	0.468	**0.480**	0.591	0.466	**0.460**
7	0.401	0.544	0.761	0.647	**0.617**	0.532	0.611	0.311	0.298	**0.269**
8	0.359	0.400	1.362	1.109	**0.621**	0.556	0.583	0.369	0.392	**0.321**
9	0.531	0.440	1.934	1.161	0.524	0.479	**0.581**	0.582	0.329	**0.272**
10	0.077	0.478	1.897	1.436	**0.485**	0.397	0.432	0.466	0.509	**0.418**
Mean	0.319	0.493	1.461	1.105	**0.545**	0.497	0.527	0.419	0.370	**0.333**
D−G	+0.175		−0.356		−0.048			−0.050		
95% interval	[+0.065, +0.297]		[−0.497, −0.212]		[−0.125, +0.016]			[−0.119, +0.019]		
D better	9/10		9/10		2/10			6/10		
D−(BD + L)					−0.030			+0.037		
95% interval					[−0.061, +0.003]			[+0.012, +0.060]		
D better					2/10			2/10		

Under median alignment, the DRAFT layout still helps (Table 10, intervals in panel b). DRAFT is above Brain-Diffuser in F-score by 0.020 and above the trajectory head by 0.008, each in nine scenes, and above the generator and MindEye in all ten. Brain-Diffuser + DRAFT layout is better still. It is above DRAFT in nine scenes and has a lower Chamfer distance in ten. Applied to the depth of Brain-Diffuser, the DRAFT layout beats the trajectory layout in F-score, Chamfer distance and SI-log. Under the log-affine fit the two layouts tie.

Under the log-affine fit, DRAFT has the highest F-score but ties the trajectory head and Brain-Diffuser + DRAFT layout, and the latter has the lower log-affine Chamfer distance. It is above Brain-Diffuser by 0.024 and above MindEye by 0.031, in all ten scenes. Its margins over the generator, the template and linear depth are 0.034, 0.050 and 0.019 (from unrounded means).

Without alignment, DRAFT has a lower SI-log than the trajectory head, generator, Brain-Diffuser and MindEye in all ten scenes ($p = 0.002$ each). Brain-Diffuser + DRAFT layout has a higher SI-log than DRAFT in all ten scenes ($p = 0.002$), by only 0.5. Linear depth ties DRAFT (lower in five scenes, $p = 0.29$). Image decoders have a higher $\delta{<}1.25$ (nine scenes each).

F.2 Agreement across views

We test whether a route's views agree by comparing their errors where two views see one point.

Change in 3D error between views. The pixel pairs come from ground truth alone, so every method is scored on the same pairs. The nine eight-view routes of a video give 72 views. For views t and $t + \ell$, we project each valid ground-truth point into the second camera. We keep the pair only if its projected depth is positive and the target pixel is valid. The two radial depths must also agree within $\max(0.10\,\mathrm{m}, 0.02r)$ for ground-truth radial depth r, and the two ground-truth points must lie within 0.10 m. With predicted points $\hat{p}$ and true points p of a matched pair, the change in 3D error is

$$E_\ell = \mathrm{mean}\,\|(\hat{p}_t - p_t) - (\hat{p}_{t+\ell} - p_{t+\ell})\|_2\,. \tag{11}$$

Table 10: **Once per-view depth scale is removed by alignment, the methods with a DRAFT or trajectory layout score best.** (a) Means over ten held-out scenes at $a = 1$. Alignment uses test depth (post hoc). F: F-score at 25 cm. SI-log: $\times 100$. Bold, underline: best and second best per column, ties share a mark. (b) Paired differences, 95% paired scene-bootstrap intervals, scenes favoring the first method and sign-flip p (Section B). †First minus second. BD: Brain-Diffuser.

(a)	Median-aligned		Log-affine fit		Unaligned pixels		
Method	F↑	Chamfer (m)↓	F↑	Chamfer (m)↓	SI-log↓	AbsRel↓	δ<1.25↑
Fixed template	0.569	0.331	0.575	0.319	52.8	1.150	0.281
Generator	0.548	0.340	0.592	0.316	54.9	0.851	0.265
Trajectory head	0.582	0.324	0.622	0.280	50.6	0.857	0.223
Linear depth	0.581	0.322	0.607	0.290	50.1	0.825	0.234
Brain-Diffuser	0.571	0.328	0.601	0.291	57.3	1.064	**0.291**
MindEye	0.558	0.345	0.594	0.297	59.5	1.030	0.283
BD + trajectory layout	0.598	0.298	0.622	0.271	51.1	0.865	0.219
BD + DRAFT layout	**0.608**	**0.291**	0.623	**0.268**	50.1	0.821	0.244
DRAFT	0.591	0.319	**0.625**	0.277	**49.6**	**0.813**	0.248
DRAFT − strongest other	−0.017	+0.028	+0.002	+0.008	−0.5	−0.008	−0.043
95% interval, lower	−0.022	+0.021	−0.002	+0.004	−1.3	−0.010	−0.074
95% interval, upper	−0.011	+0.034	+0.006	+0.013	+0.3	−0.007	−0.013

(b) Paired contrasts (Chamfer and SI-log: reduction, positive favors the first, blank: not reported)

Contrast	Metric	Difference [95% CI]	Scenes	p
Median-aligned				
DRAFT vs Brain-Diffuser	F	+0.020 [+0.009, +0.030]	9	0.006
DRAFT vs trajectory head	F	+0.008 [+0.004, +0.012]	9	
BD + DRAFT layout vs DRAFT	F	+0.017 [+0.011, +0.022]	9	
BD + DRAFT vs BD + trajectory layout	F	+0.010 [+0.006, +0.014]	10	0.002
	Chamfer (m)	+0.007 [+0.003, +0.010]	8	
Log-affine fit				
BD + DRAFT vs BD + trajectory layout	F	+0.001 [−0.003, +0.005]		
	Chamfer (m)†	−0.003 [−0.006, +0.001]		
DRAFT vs Brain-Diffuser	F	+0.024 [+0.017, +0.032]	10	0.002
DRAFT vs MindEye	F	+0.031 [+0.024, +0.037]	10	
DRAFT vs trajectory head	F	+0.003 [−0.002, +0.008]	6	0.24
DRAFT vs BD + DRAFT layout	F	+0.002 [−0.002, +0.006]	7	0.43
BD + DRAFT layout vs DRAFT	Chamfer (m)	+0.008 [+0.004, +0.013]	8	
Unaligned				
BD + DRAFT vs BD + trajectory layout	SI-log	+0.98 [+0.62, +1.41]	10	0.002
DRAFT vs Brain-Diffuser	SI-log	+7.68 [+7.09, +8.32]	10	0.002
DRAFT vs MindEye	SI-log	+9.87 [+9.05, +10.62]	10	0.002

At a lag of one second, matched ground-truth points lie 0.038 m apart on average. We average pairs within a view pair, then view pairs within a video, videos within a scene and scenes equally.

At $a = 1$, DRAFT has a smaller change than the generator and Brain-Diffuser in all ten scenes, by 0.213 m [0.190, 0.236] and 0.543 m [0.525, 0.562]. Linear depth with the DRAFT layout has the smallest change, and MindEye was not scored. The gain over Brain-Diffuser shrinks as the time between views grows (lag analysis in Section K). But a small change in 3D error is not a small error. An error shared by both views cancels, so a surface misplaced the same way in every view, such as a constant template (not scored), also has a small change. The change must therefore be read with the surface scores. In the routes drawn view by view (Section K), every method misplaces the walls, and no method puts a wall seen by consecutive views at one place.

Still, DRAFT has a smaller change than Brain-Diffuser with the DRAFT layout in all ten scenes at both factors, by 0.186 m [0.168, 0.205] at $a = 2$. Linear depth with the same layout has a similar or smaller change than DRAFT (0.581 against 0.627 m at $a = 2$, with overlapping intervals), so a small change is not specific to our generator. DRAFT itself is second best in three of the four columns of Table 11 and never best. For DRAFT at $a = 2$, the one-second change is 0.623 m within a route and 0.653 m across route boundaries. These means come from different view pairs, so they do not measure a boundary effect.

Fixed temporal smoothing. Plain smoothing over time does not help linear depth, a view-by-view decoder, in the same way. A fixed three-view filter (weights 0.25, 0.5 and 0.25 in log depth, within each route) changes the F-score by -0.0024 and raises its Chamfer distance by 0.0034 m [0.0027, 0.0043]. With the trajectory layout the changes are -0.0085 and 0.0159 m [0.0138, 0.0179]. Chamfer distance is worse in all ten scenes in both cases.

Table 11: **The DRAFT layout lowers the change in 3D error between views for every source at both factors, but at $a = 1$ it lowers the F-score of DRAFT and Brain-Diffuser.** Eight-view F-score and one-second change in 3D error E_1 of Equation (11) over the 72 views of each video. Ten-scene means with 95% scene-bootstrap intervals (paired for differences). Bold, underline: best and second best. Bottom: with minus without the DRAFT layout. BD: Brain-Diffuser.

	F@25cm, eight views ↑		Change in 3D error (m), 1 s lag ↓	
Prediction source	$a = 1$	$a = 2$	$a = 1$	$a = 2$
Current models				
Generator	<u>0.469</u> [0.442, 0.493]	0.456 [0.435, 0.477]	0.585 [0.561, 0.613]	1.227 [1.170, 1.292]
DRAFT	0.418 [0.391, 0.445]	<u>0.467</u> [0.435, 0.494]	<u>0.372</u> [0.347, 0.398]	<u>0.627</u> [0.577, 0.681]
Decoders				
Brain-Diffuser	**0.492** [0.467, 0.516]	0.411 [0.398, 0.427]	0.915 [0.887, 0.944]	2.404 [2.314, 2.498]
BD + DRAFT layout	0.436 [0.408, 0.463]	**0.481** [0.450, 0.509]	0.424 [0.397, 0.453]	0.813 [0.751, 0.880]
Linear depth	0.386 [0.351, 0.419]	0.458 [0.440, 0.477]	0.378 [0.355, 0.402]	0.660 [0.604, 0.721]
Linear depth + DRAFT layout	0.393 [0.364, 0.422]	0.446 [0.414, 0.473]	**0.353** [0.328, 0.380]	**0.581** [0.531, 0.637]
Paired differences, with minus without the DRAFT layout				
DRAFT	-0.050	$+0.011$	-0.213	-0.600
minus generator	[-0.078, -0.023]	[-0.024, $+0.043$]	[-0.236, -0.190]	[-0.671, -0.524]
BD + DRAFT layout	-0.056	$+0.070$	-0.490	-1.591
minus BD	[-0.095, -0.020]	[$+0.041$, $+0.097$]	[-0.509, -0.471]	[-1.672, -1.515]
Linear depth + DRAFT layout	$+0.008$	-0.012	-0.025	-0.079
minus linear depth	[-0.013, $+0.029$]	[-0.039, $+0.014$]	[-0.039, -0.011]	[-0.122, -0.043]

F.3 Pose drift during registration

Camera poses of a new recording may be noisy, so we test how pose drift changes the surfaces, at the registration step only. The predicted depth is fixed and uncalibrated ($a = 1$). Only the poses that lift predicted points and merge them on the 5 cm voxel grid drift, and ground truth keeps the clean poses. The drift is a horizontal translation (5, 10 or 20 cm), a yaw (1, 2 or 5 degrees) or both (10 cm with 2 degrees). It grows linearly from zero at the first view to its final value at the eighth, with five fixed seeds per route shared across levels and methods. Each method is compared with its own surface under clean poses.

The primary measure, fixed in advance, is the change in the F-score of DRAFT under joint drift. It is -0.026 percentage points [-0.100, $+0.039$], an interval that includes zero but does not establish equivalence. The Chamfer distance of DRAFT rises by 0.35 mm [0.02, 0.67]. The largest effect, 20 cm of translation for the generator, lowers the F-score by 0.165 percentage points and raises the Chamfer distance by 1.57 mm. Every mean change in every condition is below 0.2 percentage points in F-score, precision and recall and below 2 mm in Chamfer distance (all conditions omitted; Section K). These results do not show that upstream pose estimation or the full range-calibrated pipeline ($a = 2$) is robust, and they do not rank the two methods. Section I perturbs the camera-path inputs instead (post hoc).

G fMRI-free references, generator detail and range calibration

The range-calibration factor a is that of Section 4.3, and $a = 1$ means no calibration.

fMRI-free references. The **fixed template** is the per-pixel median ground-truth log depth of the nine training scenes of each fold. It is the same map for every held-out view, and we lift and score it like any prediction. Route-pooled, it reaches layout r 0.324, row order 0.581, left/right 0.509 and

pixel MAE 1.188 m. The **per-second median** instead takes, for each view, the median training-scene log depth at the same video second. It reaches layout r 0.241 and pixel MAE 1.254 m. Its layout r is below that of the fixed template by 0.081 [0.045, 0.114] (nine scenes). So knowing the video second does not help. DRAFT is above the per-second median in layout r in all ten scenes (Table 3). Its pixel MAE is lower by 0.128 m [0.028, 0.225] (eight scenes). We did not compute a readout from the constant and linear-trend regressors alone. The ten-scene mean layout r gain over the fixed template is $+0.144$ for DRAFT and $+0.157$ for Brain-Diffuser + DRAFT layout. It is $+0.109$ for the trajectory head and $+0.099$ for linear depth and the DREAM depth branch. Brain-Diffuser, MindEye and the generator stay at or below the template (Section 5.1).

The **wrong-scene control** asks how much of a surface score is specific to the scene. We lift the predictions made for the next held-out scene at the poses of the evaluated routes and score them against the evaluated scene. The correct-scene score minus this score is the scene-specific gain (Table 12). Both references are post hoc.

Table 12: **Wrong-scene predictions keep most of the F-score.** Upper rows: the fixed template scored as a prediction. Lower rows: predictions for the next held-out scene lifted at the evaluated poses (post hoc), and the gain of the correct scene, with its 95% paired scene-bootstrap interval and the number of scenes (of ten) with a positive gain. Ten-scene means. –: not applicable, since the template is the same for every scene. Brain-Diffuser and MindEye at their best factor, $a = 1$.

		Scored against the evaluated scene				Correct	Correct minus wrong scene	
Prediction	a	F@25cm↑	Precision↑	Recall↑	Chamfer (m)↓	F@25cm↑	F gain↑ [95% CI]	Scenes (of 10)
Fixed template								
Fixed template	1	0.391	0.399	0.405	0.524	–	–	–
	2	0.437	0.390	0.519	0.437	–	–	–
Wrong-scene control: predictions for the next held-out scene								
DRAFT	1	0.379	0.379	0.403	0.546	0.416	+0.037 [0.011, 0.063]	6
	2	0.415	0.356	0.527	0.492	0.484	+0.069 [0.042, 0.098]	10
Generator alone	1	0.444	0.402	0.518	0.467	0.477	+0.033 [0.017, 0.049]	9
	2	0.443	0.336	0.683	0.502	0.478	+0.035 [0.019, 0.052]	10
Brain-Diffuser	1	0.433	0.363	0.557	0.464	0.492	+0.059 [0.052, 0.065]	10
MindEye	1	0.439	0.375	0.553	0.451	0.494	+0.055 [0.040, 0.069]	10

What the generator adds. To test whether the generator adds more than the layout, we remove its detail (the high-frequency part) from the DRAFT depth, leaving only the layout, or replace it with the detail of the fixed template (all rows omitted; Section K). At $a = 2$, removing the detail lowers the F-score in all ten scenes. DRAFT minus the layout without detail is $+0.042$ [0.036, 0.047] in F-score and -0.049 m [-0.057, -0.040] in Chamfer distance. The template detail recovers part of this loss, but DRAFT stays ahead of it by $+0.025$ [0.020, 0.030] in F-score. Layout r changes by at most 0.006. Most of the detail's energy is one map shared by all scenes and views (share of detail energy in that map 0.551 against 0.007 in the ground truth, a difference of $+0.544$ [0.479, 0.648], post hoc), and it barely tracks the true detail (correlation 0.026 [0.021, 0.031]). The generator therefore acts mostly as a learned prior of indoor depth detail. The vertical stripes of $\boldsymbol{D}$ in Figure 2 belong to this shared map.

Guiding other decoders with a layout. The "+ path" and "+ DRAFT" columns of Table 2 combine a layout with the depth of another decoder. They apply the low-frequency replacement of Equation (4) to that depth instead of the generator's, using the trajectory layout or the DRAFT layout. The replacement keeps the fine detail of the decoder's depth and takes its broad structure from the layout. For the two planned decoders, the depth is the calibrated DA V2 depth of each Brain-Diffuser image or the linear depth prediction. It reuses the implementation, $\sigma = 4$ pixels, layouts and range calibration of DRAFT and has no fitted parameters. For these two decoders, all combined predictions were built and frozen under a written plan before any of them was scored. The other five decoders were added post hoc (Section C.2). Under the DRAFT layout at $a = 2$, the detail of Brain-Diffuser raises recall over that of our generator by 0.071. It lowers precision by 0.011. Both differences hold in all ten scenes (Table 14b). Without the DA V2 scale calibration, the F-scores

of Brain-Diffuser and MindEye fall to 0.242 and 0.258. Their Chamfer distances rise to 0.973 and 0.847 m. Their surface scores therefore depend on the depth scale learned from the training scenes.

How $a = 2$ was chosen. We fixed $a = 2$ after an earlier sweep over the historical models (before the loss correction) on these scenes. That sweep reused predictions of other outer folds, so the choice was descriptive and not nested. The current sweep and the nested selection below test it.

Range calibration ablation. Table 13 scores the layouts of Table 1 and the main methods of Table 2 with and without range calibration (the uncalibrated planned combinations and contrasts are summarized in Section K). Without calibration, the generator alone has the best F-score and Chamfer distance of the five layouts. Calibration raises the F-score of every layout except the generator, whose F-score barely changes, and it raises every pixel MAE. DRAFT lowers pixel MAE relative to the trajectory head in all ten scenes without calibration (by 0.077 m [0.046, 0.108]) and in nine with it (Table 3). Its wrong-scene gain in Figure 4f is 0.069 with calibration and 0.037 without it (Table 12).

Table 13: **Without range calibration ($a = 1$), fusion lowers the F-score and raises the Chamfer distance. Calibration raises the F-score of every layout except the generator's and lowers it for the image decoders.** C: EVC, M: camera motion, T: free-space features. Layout r and pixel MAE: 324-route means, F-score and Chamfer distance: ten-scene means, as in Table 1. Bold, underline: best and second best current method. Small lines: $a = 2$ minus $a = 1$, 95% paired scene-bootstrap interval. DRAFT minus generator: ten-scene means of per-scene differences (*: interval excludes zero). †Fixed after earlier results on these scenes. BD: Brain-Diffuser.

		Pixel MAE (m) ↓		F@25cm ↑		Chamfer (m) ↓	
Method (inputs)	Layout r ↑	a=1	a=2†	a=1	a=2†	a=1	a=2†
Generator alone (none)	0.303	1.168	1.507	0.477	0.478	0.426	0.449
				+0.002 [−0.012, +0.013]		+0.023 [−0.002, +0.047]	
Readout, 3×3 target (C, M)	0.440	**1.097**	<u>1.176</u>	0.432	<u>0.485</u>	0.478	0.417
Trajectory head (M)	0.390	1.217	1.264	0.383	0.447	0.535	0.443
Trajectory head (M, T)	0.446	1.185	1.252	0.397	0.465	0.514	0.427
				+0.068 [+0.044, +0.088]		−0.087 [−0.123, −0.050]	
Fusion head, DRAFT (C, M, T)	<u>0.480</u>	<u>1.107</u>	**1.153**	0.416	0.484	0.491	<u>0.405</u>
				+0.068 [+0.045, +0.090]		−0.086 [−0.126, −0.044]	
Brain-Diffuser	0.324	1.212	2.326	<u>0.492</u>	0.424	<u>0.389</u>	0.865
				−0.068 [−0.082, −0.057]		+0.477 [+0.420, +0.534]	
MindEye	0.270	1.209	2.119	**0.494**	0.416	**0.381**	0.744
				−0.078 [−0.098, −0.059]		+0.362 [+0.315, +0.404]	
BD + DRAFT layout	**0.493**	1.112	1.197	0.434	**0.499**	0.444	**0.377**
				+0.064 [+0.041, +0.087]		−0.067 [−0.110, −0.020]	
Fixed template	0.324	1.188	1.239	0.391	0.437	0.524	0.437
				+0.047 [+0.033, +0.060]		−0.087 [−0.104, −0.071]	
DRAFT minus generator	+0.172*	−0.055*	−0.356*	−0.061*	+0.006	+0.065*	−0.044*
Scenes with DRAFT better	10/10	7/10	10/10	0/10	4/10	0/10	8/10
Historical, before the correction (not ranked)							
Generator	0.369	1.219	1.435	0.426	0.463	0.513	0.518
				+0.037 [+0.023, +0.051]		+0.005 [−0.035, +0.046]	
DRAFT	0.486	1.118	1.178	0.409	0.462	0.505	0.433
				+0.052 [+0.032, +0.071]		−0.072 [−0.110, −0.032]	

Sweep and nested selection. Table 14 scores each method at a from 1 to 3. Chosen on the held-out scenes, the best factor is $a = 2$ for DRAFT, 1.5 for the generator alone and 1 for Brain-Diffuser and MindEye. At these factors DRAFT is not better than the generator (Table 15). Its F-score is lower by 0.016 and its Chamfer distance higher by 0.014 m, both within noise, and the generator has the higher recall in nine scenes. MindEye has a lower Chamfer distance than DRAFT, by 0.024 m. Every fMRI method at its best factor still beats the template in F-score (DRAFT by 0.042, nine

scenes). The generator loses precision as a grows, so part of the higher precision of DRAFT at $a = 2$ reflects the generator's over-calibration.

To avoid choosing a on the scored scene, we pick a for each held-out scene by the mean score of the other nine. By Chamfer distance, this picks the same factor as the sweep for every method except Brain-Diffuser + DRAFT layout. By F-score, DRAFT picks $a = 2.5$ in five folds, because $a = 2$ and 2.5 tie at 0.484. Its Chamfer distance then rises to 0.439 m. Its F-score falls below those of the generator and MindEye, and its Chamfer distance is worse than every image decoder's (Table 15b).

Brain-Diffuser + DRAFT layout at its best factor. At its best factor, this combination is ahead of DRAFT. Its F-score is higher by 0.015, in nine scenes (Tables 2 and 5). Its Chamfer distance is lower by 0.028 m [0.012, 0.042], in eight scenes. Against the unmodified decoders at $a = 1$, it is within noise. Its F-score is higher than those of Brain-Diffuser and MindEye by 0.006 [−0.009, 0.022] and 0.004 [−0.007, 0.016]. Its Chamfer distance is lower than theirs by 0.012 m [−0.012, 0.032] and 0.004 m [−0.020, 0.023], respectively. The F-score favors it in five scenes against each decoder, and the Chamfer distance in seven.

Table 14: **At each method's best factor, DRAFT surfaces are no better than those of the generator or the image decoders.** (a) F-score and Chamfer distance for a from 1 to 3 (bold: best a, chosen on the held-out scenes, paired differences in Table 15). (b) Precision and recall at $a = 2$ and at the best a, and DRAFT minus the row (–: reference, bold: best at $a = 2$, the $a = 1$ block is summarized in Section K). Ten-scene means, 95% paired scene-bootstrap intervals.

(a) Calibration-factor sweep

	Calibration factor a				
Method	1	1.5	2	2.5	3
F@25cm↑					
DRAFT	0.416	0.464	**0.484**	**0.484**	0.473
Generator alone	0.477	**0.500**	0.478	0.444	0.410
Brain-Diffuser	**0.492**	0.465	0.424	0.388	0.358
MindEye	**0.494**	0.469	0.416	0.373	0.340
Brain-Diffuser + DRAFT layout	0.434	0.484	**0.499**	0.491	0.474
Fixed template (control)	0.391	0.414	0.437	**0.442**	0.435
Chamfer (m)↓					
DRAFT	0.491	0.426	**0.405**	0.436	0.518
Generator alone	0.426	**0.391**	0.449	0.597	0.843
Brain-Diffuser	**0.389**	0.531	0.865	1.356	1.907
MindEye	**0.381**	0.472	0.744	1.186	1.757
Brain-Diffuser + DRAFT layout	0.444	**0.377**	**0.377**	0.439	0.557
Fixed template (control)	0.524	0.473	0.437	**0.424**	0.434

(b) Precision and recall at 25 cm

Method	Precision ↑	Recall ↑	ΔPrecision ↑ [95% CI]	ΔRecall ↑ [95% CI]
Range-calibrated, $a = 2$				
DRAFT	**0.417**	0.602	–	–
Generator alone	0.362	**0.741**	+0.055 [0.041, 0.071]	−0.139 [−0.166, −0.113]
Brain-Diffuser	0.317	0.670	+0.100 [0.085, 0.114]	−0.068 [−0.105, −0.034]
MindEye	0.311	0.654	+0.105 [0.086, 0.124]	−0.052 [−0.092, −0.017]
Brain-Diffuser + DRAFT layout	0.405	0.673	+0.011 [0.009, 0.014]	−0.071 [−0.087, −0.052]
Best tested factor by F-score, against DRAFT at $a = 2$				
Generator alone, $a = 1.5$	0.405	0.682	+0.012 [−0.004, 0.028]	−0.080 [−0.109, −0.051]
Brain-Diffuser, $a = 1$	*as above*		+0.005 [−0.006, 0.017]	−0.029 [−0.061, 0.004]
MindEye, $a = 1$	*as above*		−0.004 [−0.014, 0.008]	−0.018 [−0.043, 0.009]

H LOSS CORRECTION AND PENALTY ABLATION

An earlier implementation compared the two sides of the penalty of Equation (2) in different units.

Table 15: **Choosing a without the scored scene leaves DRAFT within noise of, or worse than, the generator and the image decoders.** For each held-out scene, $a \in \{1, 1.5, 2, 2.5, 3\}$ is chosen by the Chamfer distance or F-score of the other nine. (a) Chosen a, folds choosing it, and scores (bold: best over both selections). (b) DRAFT minus each method, 95% paired scene-bootstrap interval, scenes of ten favoring DRAFT. Ten-scene means.

(a) Selected factor and scores

	Selected by Chamfer distance			Selected by F-score		
Method	a (folds)	F@25cm ↑	Chamfer (m) ↓	a (folds)	F@25cm ↑	Chamfer (m) ↓
DRAFT	2 (10)	0.484	0.405	2 (5), 2.5 (5)	0.477	0.439
Generator alone	1.5 (10)	**0.500**	0.391	*as by Chamfer distance*		
Brain-Diffuser	1 (10)	0.492	0.389	*as above*		
MindEye	1 (10)	0.494	0.381	*as above*		
Brain-Diffuser + DRAFT layout	1.5 (6), 2 (4)	0.483	0.393	2 (10)	0.499	**0.377**
Fixed template	2.5 (10)	0.442	0.424	*as above*		

(b) Paired differences, DRAFT minus the method

	Selected by Chamfer distance		Selected by F-score	
Method	Δ [95% CI]	Scenes (of 10)	Δ [95% CI]	Scenes (of 10)
Δ*F@25cm*↑				
Generator alone	−0.016 [−0.036, 0.004]	3	−0.023 [−0.041, −0.003]	2
Brain-Diffuser	−0.008 [−0.027, 0.011]	4	−0.015 [−0.031, 0.001]	4
MindEye	−0.010 [−0.023, 0.004]	4	−0.017 [−0.029, −0.005]	2
Brain-Diffuser + DRAFT layout	+0.001 [−0.009, 0.011]	5	−0.021 [−0.027, −0.016]	0
Fixed template	+0.042 [0.025, 0.057]	9	+0.035 [0.020, 0.049]	9
Δ*Chamfer (m)*↓				
Generator alone	+0.014 [−0.009, 0.036]	3	+0.048 [0.010, 0.093]	2
Brain-Diffuser	+0.016 [−0.006, 0.039]	3	+0.050 [0.023, 0.086]	0
MindEye	+0.024 [0.008, 0.042]	1	+0.058 [0.031, 0.096]	0
Brain-Diffuser + DRAFT layout	+0.012 [0.001, 0.022]	3	+0.062 [0.047, 0.078]	0
Fixed template	−0.019 [−0.040, 0.002]	7	+0.015 [−0.012, 0.045]	4

The correction. The penalty reads the layout code of a generated depth map q through a fixed affine inverse of the renderer, $\mathcal{R}^{+}_{\text{aff}}(q) = y_* + J^{+}\,\text{vec}(q - q_*)$ with $q_* = \mathcal{R}(y_*)$. Here y_* is the mean training layout code and J a central-difference Jacobian of $\mathcal{R}$ at y_*, mean eye height and zero pitch. This inverse returns the code in its original units, but the posterior is standardized. The correction standardizes the inverse like the posterior mean of each sampled recording, giving the map $\mathcal{A}_i$ of Section A.2. We call checkpoints trained before this change *historical* and the retrained ones *corrected*. The historical generator without brain input received a zero code, which need not render as the mean training layout.

Effect on the results. All ten folds were retrained with everything else unchanged. Predictions were frozen before scoring, and no checkpoint or factor was chosen with the new scores. The corrected generator has a lower layout r (0.303 against 0.369) but a lower pixel MAE and better surfaces at $a = 1$. The DRAFT layout stays close to its historical value (0.480 against 0.486). Table 13 lists the historical and corrected rows side by side. Range calibration raises every pixel MAE. The main paper uses the corrected models throughout. An independent replay reproduced surface scores to within 4.75×10^{-5} in F-score and 2.71×10^{-6} m in Chamfer distance.

H.1 Penalty ablation

To test whether the penalty of Equation (2) helps, we retrained the corrected generators of all ten folds without it ($\lambda = 0$) and compared them with $\lambda = 0.1$. Both versions share seed, data, architecture and evaluation. The primary measure, fixed in advance, was the Chamfer distance of the generator at $a = 1$. Without the penalty it rose by 0.1032 m [0.0710, 0.1399], in all ten scenes, and

its F-score changed by -0.065 [-0.087, -0.045]. The surfaces of DRAFT also worsened, with F-score changes of -0.021 ($a = 1$) and -0.022 ($a = 2$) in all ten scenes. The generator's F-score at $a = 2$ changed within noise.

The penalty does not help every measure. Without it, the generator's layout r rose by 0.1172 (from 0.301 to 0.418). Its range ratio, the predicted over the true spread of cell log depths (Section C.1), fell from 0.7835 to 0.6501 at $a = 1$, farther from the ideal of one, and from 1.567 to 1.300 at $a = 2$, closer to one. Its pixel MAE at $a = 2$, which scores the depth map before rendering, fell by 0.283 m. The penalty thus trades the generator's layout r for better surfaces. This result rests on one training seed (all 18 measures with intervals omitted; Section K).

I Post hoc robustness analyses

All analyses in this section are post hoc. We designed them after every result above was known and ran them on the same ten scenes. Before running any of them, we wrote down six primary tests with their decision rules and all secondary quantities. Only ridge heads and ridge readouts were refit, so the generator, the readout inputs and every number elsewhere in the paper are unchanged. Independent code recomputed each primary result, and we report every outcome, favorable or not. Statistics follow Section B (ten scenes as units, 95% paired scene-bootstrap intervals with 20,000 resamples, exact sign-flip tests over 1,024 sign patterns). The six primary tests are Holm-adjusted together (Table 16). Steps added after a result was seen are marked. Section K summarizes all secondary intervals.

What we had seen before the plan was frozen. Two results revisited here were known before we froze the plan. The first is the single-recording gain of DRAFT over the trajectory head in layout r ($+0.016$, eight scenes). The second is the surface scores of DRAFT at four distance thresholds for one scene. No primary contrast had been seen, including feature replacement with single recordings.

Table 16: **Post hoc primary tests: two support a hypothesis, three are null and one goes against it.** Estimate: ten-scene mean and 95% paired scene-bootstrap interval. Scenes: scenes of ten with a positive value. p: exact sign-flip. Holm: adjusted over the six tests. Outcome: the decision rule written before the analysis. *A time-misaligned brain component gives the same slope, so this test does not show brain signal (text). †The hypothesis predicted a negative value.

Question	Statistic	Estimate [95% CI]	Scenes	p	Holm	Outcome
Does the brain's gain survive single recordings?	Layout r, DRAFT minus feature replacement, one recording, no pooling	$+0.012$ [0.005, 0.018]	8	0.010	0.049	support
Does averaging more recordings raise layout r?	Slope per doubling of recordings (1, 2, 3, 6)	$+0.0047$ [0.0038, 0.0055]	10	0.002	0.012	support*
Is the gain larger where the path covers less of the view?	Within-scene Spearman of path coverage and per-view gain	$+0.042$ [-0.021, 0.110]	7	0.27	0.82	null†
Does the gain over the path grow with path noise?	Slope of DRAFT minus trajectory head over six noise levels	-0.0006 [-0.0047, 0.0036]	4	0.81	0.82	null
Is DRAFT more scene-specific than MindEye?	Correct minus wrong-scene F@25cm, DRAFT minus unmodified MindEye	$+0.014$ [-0.009, 0.042]	6	0.37	0.82	null
Does DRAFT gain on MindEye at coarser thresholds?	Slope of the F-score difference per doubling of the threshold (10 cm to 1 m)	-0.0053 [-0.0079, -0.0024]	2	0.010	0.049	contrary

Single recordings without pooling. By a protocol written before this plan, we decoded each of the 416 recordings alone, with no averaging over recordings and no pooling across styles, using the models fitted on the other nine scenes. Both versions use the generator output conditioned on that recording, so only the layout differs. Replacing the EVC features with their training mean lowers the layout r of DRAFT by 0.012 (primary test), in eight scenes and 277 of 416 recordings. This is about a third of the pooled effect with the control layout on the same generator (0.034). Scenes 4 and

6 are slightly negative (-0.002 and -0.003). A circular BOLD shift lowers layout r by 0.016, also in eight scenes. Single-recording DRAFT is below pooled DRAFT in every scene (0.454 against 0.473). Its gain over the trajectory head is 0.016 in eight scenes, a contrast seen before the plan that scores the two with different generators. With both on one generator, the gain is 0.018 in nine scenes, and 17 of 20 participants have a positive mean gain (descriptive, as participants are nested in scenes). The readout with a 3×3 target does not beat its motion-alone version on single recordings (-0.005, four scenes), so its advantage in Table 3 appears only with averaged recordings. The F-score of single-recording DRAFT is not lower than that of pooled DRAFT ($+0.008$, six scenes). These are single recordings at test time, not single-participant models, because the fitted models include the same participants' other scenes.

More recordings of the same videos. Averaging $k = 1$, 2, 3 or 6 recordings per video before fusion (up to 30 random subsets per k) raises layout r from 0.456 to 0.463, 0.465 and 0.468. This slope of 0.0047 per doubling holds in all ten scenes and meets the support rule. In a check added after the result, a 37 s circular shift of each recording's brain component, breaking its alignment with the video, left the slope unchanged (difference -0.00003, five scenes). The rise thus shows only that single recordings are noisy, and the brain signal is in the levels. The gain over feature replacement is 0.013 at $k = 1$ (eight scenes), 0.025 at $k = 6$ and 0.026 with all recordings (ten scenes each), and at $k = 6$ the intact component beats the shifted one by 0.036 in all ten scenes. A fit of $c_0 - c_1/k$ gives an infinite-recording layout r of 0.470. All recordings reach 0.468, and pooled DRAFT (0.473) is at this limit (-0.003), so more recordings of the same videos would add about 0.002. The extrapolated gain over the trajectory head, 0.032, is that reported in Section 5.2. The brain component is reproducible across participants (split-half correlation within a video 0.61). For the readout with a 3×3 target, the extrapolated limit is 0.438, equal to the trajectory head ($+0.000$). These limits describe this readout and fusion head, both trained on averaged recordings, not the information available in EVC.

Fixed design settings and acquisition blocks. To test whether choosing the fixed settings on these scenes created the brain's gain, we refitted the trajectory and fusion heads, with the same readout inputs and generator, for 300 combinations of the trajectory window K (4, 8, 16, 37 and 73 s), the trajectory penalty (10 to 10^4), the fusion penalty (1 to 10^4) and the low-pass width σ (2, 4 and 8 pixels) (Section K). DRAFT beats the trajectory head in all 300 settings, by 0.010 to 0.064 (median 0.039), in at least eight scenes in 293 settings. The other seven settings have seven positive scenes, and six of them combine $K = 73$ s with a penalty of 10^4. The gain over feature replacement ranges from 0.030 to 0.052 (median 0.041), in at least nine scenes in all 300 settings. The paper's setting does not favor the brain, because its gains (0.035 and 0.033) are at or below these medians. It does favor both heads' absolute scores. Only 16 of the 300 settings reach its DRAFT layout r of 0.480 (median 0.457, range 0.422 to 0.488), and the trajectory head is also near its best (0.446 against a median of 0.415). The paper's absolute layout r values are therefore optimistic by 0.02 to 0.03 for both heads, but their difference is not. DRAFT stays above linear depth alone (0.429, not re-tuned) in 273 of 300 settings, and all 27 failures have $\sigma = 8$. Split into the three acquisition blocks (scenes 1 to 4, 5 to 7 and 8 to 10), both brain-input controls and the gain over the path are positive in every block. This follows from their being positive in all ten scenes, so it is not separate evidence, and the blocks share most participants. The one block contrast that could fail, DRAFT minus linear depth alone, is $+0.094$ and $+0.035$ in the first two blocks but -0.011 in scenes 8 to 10 (one of three scenes). Refitting both heads without the held-out block keeps the gain over the trajectory head at 0.041, in all ten scenes (readout and generator not refitted). Over 55 settings of voxel count, penalty, lag window and ensembles, the readout with a 3×3 target stays above its motion-alone version by 0.079 to 0.102 in at least nine scenes, but it is never above the trajectory head (best 0.442).

Where the brain's gain arises. To test whether the brain helps most where the camera path says little, we define the path coverage of a view as the number of its 12 azimuth bins of 7.5° that the camera visits within ±4 s, the window of the trajectory head. The primary test found no negative within-scene association between coverage and the per-view gain of DRAFT over feature replacement (mean Spearman correlation $+0.042$, negative in only three scenes). Descriptively, the gain is 0.013 in views whose path covers no in-view bin (five scenes positive) and 0.043 in the others (ten scenes). The gain does not fill in where the path is blind.

Noisy camera paths. At test time we added independent per-second jitter to horizontal position and heading, at five levels from 5 cm and 1° to 1 m and 20° (10 seeds each). For scale, the clean camera moves 0.33 m and turns 14° per second on average. The jitter enters the trajectory features, the readout's camera-motion regressors and the fusion head, all trained on clean paths. Pooling partners and masks come from the clean path. The primary test, that the gain over the trajectory head grows with noise, was not supported (slope -0.0006, $p = 0.81$). The gain rises from 0.035 to 0.063 at 0.5 m and 10° (ten scenes at each level) and falls to 0.013 at 1 m and 20° (seven scenes). DRAFT itself is not robust. Its layout r falls from 0.473 to 0.468, 0.460, 0.419, 0.284 and 0.083, about as fast as that of the trajectory head (0.438 to 0.070). From 0.5 m and 10° on, DRAFT is below the readout with a 3×3 target, which keeps 0.393 and 0.240 (differences -0.109 and -0.157, no scene favoring DRAFT), and below a version of itself without the path input (0.405). The fusion head relies on the path, so a path this noisy is worse than none. The gain over feature replacement stays positive at every level (0.025 to 0.053, ten scenes each), as expected, because the brain term is the same at every level. The generator detail was not resampled. Section K summarizes these results and the checks with re-matched pooling partners, cumulative drift and heads refitted on noisy paths.

Scene-specific surface scores and distance thresholds. To test whether removing the wrong-scene score of Table 12 changes the surface comparison, we reproduced every archived score within 10^{-6} and scored each method against its wrong-scene predictions (Section G) at 10, 25, 50 and 100 cm. The scene-specific gain, correct minus wrong-scene F-score, is 0.069 for DRAFT and 0.055 for MindEye. Their difference, $+0.014$ in six scenes, comes mostly from scenes 4 and 6, and the primary test is null. It also depends on range calibration. Without it, the gain of DRAFT is below that of MindEye (-0.018, four scenes, added after the result). The F-score of DRAFT minus that of MindEye is $+0.005$ at 10 cm (seven scenes), -0.010 at 25 cm (four), -0.020 at 50 cm (one) and -0.011 at 1 m (three). The primary test, that DRAFT gains on MindEye at coarser thresholds, therefore went against the hypothesis (slope -0.0053 per doubling, Holm-adjusted $p = 0.049$), with the deficit appearing between 10 and 25 cm. Against Brain-Diffuser, the differences are $+0.006$, -0.008, -0.014 and -0.007, and no difference in scene-specific gain exceeds noise. Within our pipeline, the layout makes the surfaces more scene-specific than the generator alone calibrated in the same way ($+0.034$ at 25 cm, eight scenes). Of the methods scored here, Brain-Diffuser + DRAFT layout has the highest mean F-score at 25 cm, 50 cm and 1 m (0.499, 0.757 and 0.930), against 0.494, 0.752 and 0.927 for MindEye, a difference within noise at 25 cm (Section G). No wrong-scene-corrected or threshold score shows the surfaces of DRAFT beating MindEye beyond noise.

What these analyses do not show. They reuse the development scenes, participants and scanner runs, so they test robustness, not generalization to new scenes or people. One of the two supported hypotheses, the rise with more recordings, does not show brain signal. The single-recording result has Holm-adjusted $p = 0.049$, close to the threshold. The design grid varies four fixed settings. The generator, pooling rule, reliability masks, 3×3 cells, fusion architecture and 22-number layout code stay as chosen in development. The path noise is synthetic, and the limit over recordings is extrapolated.

J Post hoc analyses: withheld scenes and low-level image features

We added these two analyses last, after every result above was known. Each has a written plan with decision rules, time-stamped before any of its outcomes was computed, and each was recomputed with independent code. We report every outcome here or in Section K and mark every step added after a result was seen. No other number of the paper changes. Section J.2 reuses the ten development scenes with the statistics of Section B. Section J.1 uses scenes withheld from fitting, selecting and tuning DRAFT.

J.1 Scenes withheld from development

The first withheld scene and its prior exposure. The first withheld scene (W1), one more of the same design (one floor plan in four styles), was recorded in the session of scenes 5 to 7 (75 recordings from 19 of the 20 participants). No component of DRAFT was fitted, selected or tuned

on it, and every script that produced DRAFT refused its data. It is not unseen, though. Before DRAFT was designed, we tested an earlier readout on it once, and the recorded decision was a pass. That readout used the same six ridge configurations, trained with other scenes, but predicted a different target (13 cosine coefficients of inverse depth). The layout code, pooling, generator and both heads were designed after that outcome, and earlier generations of our pipeline had also been run on the scene. In one style the rendered camera shakes around the recorded path, so ground truth ray-cast along that path does not match what participants saw, and the plan excluded that style from the primary analysis. Five further scenes were set aside during development. One has no validated camera path and stays excluded. The other four are W2 to W5.

Plan for W1. The plan was frozen before any prediction for W1 existed. Every component of DRAFT was retrained on all ten scenes with the paper's settings. Inner selections kept the values chosen in most development folds (a 4 s trajectory window and penalties of 10^3 for both heads), the generator was retrained for 20,000 steps, and all fields were frozen with hashes before the ground truth was rendered. Before any data of W1 were loaded, a rehearsal with two development scenes held out reproduced their published predictions within 10^{-9}. We score 27 eight-view routes in the three primary styles (56 recordings). Three primary contrasts in layout r compare DRAFT with the trajectory head (H1), the fixed template (H2) and feature replacement with the generator resampled as in the paper (H3). With one scene, the unit of inference is the participant. A cluster bootstrap over the 19 participants (5,000 resamples) recomputes pooling, posterior, fusion and generator detail, and one-sided p values are Holm-adjusted over the three contrasts. A contrast replicates if it is positive with adjusted $p < 0.05$. Route and style counts are descriptive, because routes share a floor plan and, within a video, a camera path.

Results in W1. All three primary contrasts met the replication rule (W1 row of Table 17). Absolute accuracy did not transfer, though. DRAFT reached a layout r of 0.285, below every development scene and below the 95% prediction interval from the ten scenes (0.299 to 0.647). Every other layout was also lower than in any development scene, including the fixed template (0.178) and the generator alone (0.149), so the scene is harder for all methods. Feature replacement lowered layout r in 22 of 27 routes and in every style. The gain over the trajectory head, 0.013, is about 40% of its development mean, positive in only 16 of 27 routes and negative in one style. The planned bootstrap varies only the participants, so it is narrow. Resampling routes instead, a descriptive step added after the result, gives intervals of -0.012 to $+0.039$ for H1, $+0.017$ to $+0.204$ for H2 and $+0.013$ to $+0.049$ for H3. The surfaces ($a = 2$, leave-one-participant-out jackknife intervals) do not reproduce the development ordering in F-score. DRAFT has a lower F-score than the generator alone calibrated in the same way (0.442 against 0.456, -0.014 [-0.027, -0.001]) but a lower Chamfer distance (-0.047 m [-0.061, -0.033]). Its F-score is above those of the trajectory head, feature replacement, the circular shift and the fixed template ($+0.022$, $+0.033$, $+0.032$ and $+0.068$), and its Chamfer distance is below all four. Section K summarizes every arm and style of W1.

The shaking-camera style. Decoded alone without pooling across styles (9 routes, 19 recordings), DRAFT was below the trajectory head in this style (-0.026 [-0.033, -0.018]), a contrary result by the plan's rule. Pooling all four styles by the rule of the paper (36 routes, 75 recordings) replicated all three primary contrasts, with $+0.013$ [0.004, 0.021] for H1 and $+0.032$ [0.023, 0.040] for H3.

Four more withheld scenes. We analyzed W2 to W5 after the results of W1 were known, but their plan was frozen with that of W1, before any prediction for W1 existed. They have the same design and were recorded in a later session from six participants, all of whom also appear in the development scenes and five of whom also appear in W1. W2, W4 and W5 have four styles with six recordings each, and W3 has one style (the only one with ground truth) and six recordings, for 78 recordings and 117 eight-view routes. No camera shakes in these videos, so every style was scored. These scenes are more exposed than W1. They were training scenes of earlier model generations in this project, the earlier readout tested once on W1 had been trained with them, and on 2026-09-15, before W1 was tested, we had confirmed that this readout works on them. As for W1, no component of DRAFT was fitted, selected or tuned on them, and every script that produced DRAFT refused their data.

Procedure for W2 to W5. We applied the models of W1 (generator, fusion stack, fixed template and both heads) without refitting. Before any data of W2 to W5 were loaded, a rehearsal reproduced every prediction for W1 bit for bit, including the generator samples. We then froze the predictions of every arm with hashes and a time stamp and read the first ground-truth depth only after the freeze. This ground truth was already stored with the data, so we enforced the order by blanking those rows and logging every file that the analysis processes opened. The logs cannot rule out a read by a separate process, but after the rehearsal every prediction was a deterministic, seeded function of the frozen models, so a look at the ground truth could change no choice. Scoring followed W1, with the bootstrap over six participants. The planned five-scene summary is the mean of the per-scene differences with an exact one-sided sign-flip p over the 32 sign patterns. One deviation was written down before any data of these scenes were read. W3 has one recording per participant, so we waived a rule that requires two (it only supplies the partner recording for a control not used here) and also report a four-scene summary without W3. W3 has only one video, so pooling has no partner there. An independent reviewer reproduced every number with separate code (within 10^{-15}) and found no departure from the plan.

Table 17: **Five withheld scenes: DRAFT beats the fixed template in every scene, but the brain's contribution replicates within only two.** Layout r, route-pooled. Differences: DRAFT minus the named layout, with 95% participant cluster bootstrap intervals that omit variation between routes. *Replicated (positive, Holm-adjusted $p < 0.05$ over the three contrasts of the scene). Mean: mean of the five differences, positive scenes and exact one-sided sign-flip p (1/32 is the smallest attainable value, and the scenes are not independent, see text). Development: ten-scene mean and range of the differences (layout r of DRAFT 0.359 to 0.578).

Scene	Recordings (participants)	DRAFT	H1: trajectory head	H2: fixed template	H3: feature replacement
W1	56 (19)	0.285	+0.013 [0.006, 0.020]*	+0.107 [0.099, 0.114]*	+0.030 [0.022, 0.036]*
W2	24 (6)	0.489	+0.031 [0.003, 0.045]*	+0.231 [0.202, 0.246]*	+0.078 [0.048, 0.092]*
W3	6 (6)	0.476	+0.004 [−0.027, 0.030]	+0.211 [0.181, 0.236]*	+0.034 [0.001, 0.061]*
W4	24 (6)	0.579	+0.033 [−0.001, 0.044]	+0.197 [0.163, 0.209]*	+0.025 [−0.009, 0.037]
W5	24 (6)	0.534	+0.013 [−0.029, 0.035]	+0.195 [0.152, 0.219]*	+0.041 [−0.004, 0.065]
Mean, scenes, p			+0.019, 5, 0.031	+0.188, 5, 0.031	+0.042, 5, 0.031
Development		0.473	+0.035 (0.009, 0.064)	+0.144 (−0.012, 0.307)	+0.037 (0.012, 0.075)

Results in each scene. Absolute accuracy was in the development range in the four later scenes (Table 17). DRAFT reached a layout r of 0.489, 0.476, 0.579 and 0.534 in W2 to W5, all inside the 95% prediction interval from the ten scenes (0.299 to 0.647), with W4 marginally above the top of the development range (0.578). Every layout was higher than in W1, so W1 was the hardest scene for every method. H2 replicated in all five scenes. H1 replicated only in W1 and W2 (W2: adjusted $p = 0.016$) and was positive but missed the rule in W3 ($p = 0.43$), W4 ($p = 0.060$) and W5 ($p = 0.33$). H3 replicated in W1, W2 and W3 (W3: adjusted $p = 0.046$) and was positive but missed the rule in W4 ($p = 0.083$) and W5 ($p = 0.072$). The circular shift lowered layout r in all four later scenes ($+0.077$, $+0.038$, $+0.029$ and $+0.040$), with intervals that exclude zero only in W2 and W3. Every primary difference in every scene fell inside its development prediction interval. On single recordings without pooling, DRAFT beat the trajectory head in 10 of 24, 2 of 6, 10 of 24 and 11 of 24 recordings of W2 to W5 (W1: 31 of 56), so the gain over the path appears only when recordings are pooled across participants. Section K summarizes route counts and two secondary results of W5 that went against the development pattern.

Summary over the five scenes, and the unit of inference. DRAFT was above the trajectory head, the fixed template and feature replacement in all five scenes, with mean differences of $+0.019$, $+0.188$ and $+0.042$ and an exact sign-flip p of $1/32 = 0.031$ each (Table 17). This is the smallest p that five scenes allow, so it says only that all five differences were positive, not how large they are. W3 counts as positive for H1 with $+0.004$ and a within-scene p of 0.43. The plan set no significance level or multiplicity rule for this summary, and Holm adjustment over the three contrasts gives 0.094 for each. Without W3, the means are $+0.023$, $+0.182$ and $+0.043$, with four of four scenes positive and $p = 1/16$ each. The circular shift gives $+0.044$ (five scenes, $p = 1/32$), and the readout with a 3×3 target minus its motion-alone version gives $+0.105$ (four scenes, $p = 0.063$). The sign-

flip test treats the five scenes as independent, and they are not. W2 to W5 share six participants and a session, five of the six also appear in W1, and one frozen model scores all five scenes. This positive dependence makes $p = 1/32$ too small by an unknown amount. Within each scene the unit is the participant, and with six participants the intervals of W2 to W5 are rough. By the plan, no combination of these scenes is a test of generalization across scenes.

Surfaces. At the paper's factor $a = 2$, DRAFT had a higher F-score than the generator alone in W2, W4 and W5, by $+0.055$ [0.020, 0.089], $+0.047$ [0.016, 0.079] and $+0.070$ [0.056, 0.084] (jackknife intervals). It tied the generator in W3 (0.000 [-0.035, 0.035]) and was lower in W1 (-0.014 [-0.027, -0.001]). Its Chamfer distance was lower than the generator's in all five scenes, by 0.085 m to 0.128 m in W2 to W5. Against the trajectory head, the F-score was higher in W2 ($+0.063$ [0.038, 0.088]) and W5 ($+0.038$ [0.022, 0.054]), within noise in W4 ($+0.024$ [-0.006, 0.055]) and lower but within noise in W3 (-0.013 [-0.047, 0.020]), where the Chamfer distance of DRAFT was worse ($+0.058$ m [0.012, 0.104]). Without range calibration ($a = 1$) the ordering reverses: DRAFT has a lower F-score than the generator alone in all five scenes and a worse Chamfer distance (point values). The F-scores are 0.375 against 0.439 in W1, 0.424 against 0.510 in W2, 0.360 against 0.472 in W3, 0.465 against 0.529 in W4 and 0.423 against 0.484 in W5. The surface gain over the generator thus depends on the range calibration, whose factor was fixed on the development scenes.

What the withheld scenes do and do not show. With every prediction frozen before the ground truth was read, DRAFT beat the fixed template within all five withheld scenes and was above the trajectory head and feature replacement in all five. The brain's contribution as defined in advance, the replication of both H1 and H3 within one scene, held in only two (W1 and W2). The gain over the trajectory head averaged $+0.019$, about half its development mean ($+0.035$), varied between routes and styles, reversed in the shaking-camera style of W1 and appeared only after pooling across participants. The absolute layout r transferred in the four later scenes but not in W1. The surface gains over the generator alone held at $a = 2$ in three of the four later scenes and reversed at $a = 1$ in all five. These scenes were withheld, not unseen, and the earlier readout was known to work in W2 to W5. They share participants, one model and, for W1, the scanner session of scenes 5 to 7, so they are not independent replications and say nothing about new participants.

J.2 Low-level image features

We test whether EVC carries layout information beyond the low-level content of the frames. For each 3×3 cell of the frames shown at second t, we computed mean luminance, root-mean-square contrast, log Gabor energy (3 scales, 4 orientations) and log frame-difference motion energy within ±0.5 s. Principal component analysis on the training scenes reduced these 135 features to 16 components. We refitted the readout with a 3×3 target of Table 3 with four input sets and all else as in the paper (six ridge configurations with fixed penalties, leave-one-scene-out folds, and the same pooling, low-frequency replacement and generator). The sets are the 12 camera-motion regressors alone (M0), motion plus the image components (M1), motion plus the 16 EVC components (M2, the paper's readout, reproduced within 10^{-6}) and motion plus both (M3). The primary test compares M3 with M1. Earlier exploratory analyses pointed to low-level mediation, so its direction was not unexpected.

Table 18: **EVC adds no measurable layout r to low-level image features of the frames.** Readout with a 3×3 target, layout r over 324 routes. Differences: ten-scene mean [95% paired scene-bootstrap CI], scenes of ten with a positive value, exact sign-flip p. Secondary tests are Holm-adjusted over four. Left: readouts. Right: contrasts, not tied to the row beside them.

Inputs	Layout $r\uparrow$	Contrast	Difference [95% CI]	Scenes	p	Role
M0: camera motion	0.338	M3 − M1	$+0.013$ [-0.006, 0.033]	6	0.25	primary: null
M1: + image features	0.561	M1 − M0	$+0.223$ [0.203, 0.244]	10	0.008 (Holm)	secondary
M2: + EVC (paper)	0.440	M2 − M0	$+0.098$ [0.053, 0.152]	9	0.008 (Holm)	secondary
M3: + image features + EVC	0.573	M3 − M2	$+0.138$ [0.109, 0.164]	10	0.008 (Holm)	secondary
		M2 − M1	-0.125 [-0.161, -0.083]	1	0.008 (Holm)	secondary

The primary test is null (Table 18). Adding EVC to motion and image features changes layout r by $+0.013$ [-0.006, 0.033], positive in six scenes, $p = 0.25$. Every planned sensitivity analysis agrees: $+0.013$ and $+0.012$ with 8 or 32 image components, $+0.006$ with all 135 features, $+0.010$ with inner-validated penalties, and $+0.019$ [-0.003, 0.042] with the image features passed through a canonical hemodynamic response and the lag window of EVC. No p is below 0.15, and the layout without generator detail gives $+0.014$ ($p = 0.21$). With ten scenes, an EVC-specific gain of up to about 0.03 cannot be excluded. An exploratory measure points the other way. EVC explains unique variance in the per-view shape of the 3×3 layout before the generator (unique cross-validated R^2 $+0.042$ [0.016, 0.069], seven scenes, $p = 0.018$). M1 beats the paper's readout (0.561 against 0.440), partly because the image features are exact for second t and BOLD lags by seconds. With hemodynamically filtered features, the model without brain data reaches 0.474, within noise of 0.440 ($+0.034$ [-0.005, 0.070], eight scenes, $p = 0.12$, added after the result). Like depth estimated from the video frames (Table 2), M1 uses the stimulus frames, which DRAFT never sees, so it is a stimulus-side reference, not a competing decoder. In this dataset, the layout information of EVC is not measurably separable from the low-level content of the frames. The features do not cover eye movements or other low-level features, and the limitation in Section 6 stands.

K Additional analyses

The analyses below were run on the same data but are not included in this version for space. Each line names one and states its outcome. The interactive viewer in the supplementary material shows all ten atlas routes.

- **Re-rendered depth metrics:** definitions of the re-rendered depth metrics used only by the DREAM comparison below.
- **Image-decoder outputs:** decoded images and depth of the image decoders on the atlas views. DRAFT has the lowest mean pixel MAE on these views, but over all routes it is worse than linear depth in scenes 1 and 8 and better in scene 4.
- **DREAM depth branch, with its endpoint definitions:** the DREAM depth branch has a better layout r than our generator but worse surfaces (F-score lower in all ten scenes).
- **All transfer scores:** at $a = 1$ the DRAFT layout lowers the F-score of Brain-Diffuser, MindEye and CLIP retrieval.
- **All rows of the generator-detail ablation:** at $a = 1$, removing the detail lowers the F-score by 0.031, and the template detail recovers part of this. At $a = 2$ the detail raises pixel MAE by 0.016 m.
- **Uncalibrated planned combinations and contrasts:** the planned combinations reach F-scores of 0.371 to 0.434, below the calibrated full DRAFT (a comparison including the calibration), and the two with the DRAFT layout have a slightly higher layout r than DRAFT. Uncalibrated DRAFT has a lower pixel MAE than the generator and trajectory head but a lower recall than the generator, with no scene favoring DRAFT.
- **Precision and recall at $a = 1$:** uncalibrated, DRAFT has a lower recall than the generator, Brain-Diffuser and MindEye, with precision within noise.
- **Unaligned F-score thresholds:** no unaligned F-score threshold from 10 to 50 cm separates DRAFT ($a = 2$) from Brain-Diffuser or the generator ($a = 1.5$). Within noise, image decoders have higher F-scores at 35 and 50 cm.
- **View-by-view lifting:** three routes lifted view by view. By the change in 3D error, DRAFT views agree more than those of Brain-Diffuser + DRAFT layout, but the two look alike by eye.
- **Duration and lag:** the F-score of DRAFT rises from 0.445 over 6 views to 0.595 over 30, so no best duration can be identified. The change in 3D error grows with lag while the methods converge. At eight views DRAFT has a lower recall than the generator.
- **Historical rows:** before the loss correction, the generator had a lower eight-view F-score at $a = 1$ (0.420 against 0.469) but its views agreed more (one-second change 0.441 against 0.585 m), and the change of DRAFT was the same (0.372 m).
- **Registration drift, all conditions:** registration drifts shift means by < 0.2 percentage points and 2 mm.
- **Per-scene neural controls:** aligned brain input gives a higher layout r and F-score in all ten scenes under both neural controls.
- **Equal-scene-weight matched cohorts:** an equal-scene-weight variant of the matched cohorts gives a similar result.
- **Penalty ablation:** without the penalty DRAFT has a slightly higher layout r and a slightly lower pixel MAE, with the same range ratio.
- **Per-scene uncalibrated ($a = 1$) rows:** DRAFT has a lower F-score and a higher Chamfer distance than the generator in all ten scenes.
- **Surface atlas of the ten atlas routes:** failures included, with the scores of Table 9. DRAFT has the lowest route F-score of three methods in seven of the ten routes.

- **Depth atlas of scenes 5 to 10:** DRAFT has the highest route layout r of three methods in scenes 5, 6 and 9, and Brain-Diffuser + DRAFT layout in scenes 7, 8 and 10.
- **Noisy-path checks:** with re-matched pooling partners the gain over feature replacement stays positive. Under cumulative drift DRAFT falls from 0.473 to 0.412 while its gain over the trajectory head grows. Heads refitted on equally noisy paths recover most of the loss, with a gain (0.055) no larger than under test-time noise, and the refit left the training brain inputs unperturbed.
- **Design settings and secondary intervals:** the 300 design settings; intervals of the 24 secondary contrasts of Section I.
- **Secondary results:** a fit on $k \leq 6$ recordings predicts the held-out $k = 12$ value within 0.0002. Much of the brain component is shared by videos with the same camera path (added after the result). Against the hypothesis, path coverage is positively associated with the gain over the circular shift and over a layout-only version (unadjusted $p = 0.035$ and 0.010, neither surviving correction). Five other path measures, the gain over the trajectory head and a per-cell version show no detectable association. A second per-cell version, added after the result, points toward the hypothesis (unadjusted $p = 0.047$).
- **Withheld scene W1:** every arm and style of W1. The readout's gain in code similarity is slightly smaller (0.192 against 0.219), and DRAFT has a higher precision but a lower recall than the generator.
- **Shaking-camera style of W1:** DRAFT beats feature replacement and the fixed template there, but not the trajectory head.
- **W2 to W5:** route counts, and two secondary results of W5 against the development pattern (the direct readout below its camera-only version, and a negative gain in code similarity).
- **Layout-code ablation (post hoc):** by the planned rule the wall distances, floor extents and opening shares each carry the brain's gain, the median and spread of depth do not, and no group is necessary. The gain raises the left-right correlation across views ($+0.039$, unadjusted) but not row order.