PRISM: PREDICTING COMPLETE SCENE-MEMORY REPORT DISTRIBUTIONS FROM VISUAL GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual working memory preserves information about a scene after it is no longer visible. Yet it is usually modeled through accuracy or error magnitude, which discard which scene a person reports. Two equally distant errors can differ in visual content, making error identity a richer probe of remembered representations. We introduce PRISM, a framework for predicting complete distributions of scene reports from target–candidate visual similarity. PRISM-Eval isolates what frozen vision networks contribute beyond response-display structure and six fixed image descriptors. PRISM-Channel places visual evidence within a cognitive model of memory precision and guessing, while Cognitive-Channel Geometry Distillation (CCGD) learns a compact combination of frozen similarities. Across two continuous-report datasets, intermediate visual features improve held-out likelihood beyond all fixed controls. Combining three visual models improves cognitive-baseline predictions, with relative-weight gains under fixed regularization and positive adapted transfer to an independent memory task. Independent EEG data show that the behaviorally useful geometry captures neural structure beyond six fixed descriptors. PRISM thus turns memory error identity into a controlled bridge between natural-scene vision, working-memory theory, and neural representations.

1 INTRODUCTION

Visual working memory maintains scene information after stimulus offset. Two incorrect choices can be equally far from the target in a response display yet differ in how often people select them. Error identity—not only error magnitude—therefore reveals structure in the maintained representation (Brady et al., 2009; Taylor & Bays, 2018; Bays et al., 2024).

PRISM (Predictive Representations In Scene Memory) predicts memory reports from target–candidate visual similarity. In CCGD, training reports supervise similarity weights; new reports are outcomes, never inputs. Here *visual geometry* means pairwise image similarities, not the response display's layout. PRISM-Eval assigns probabilities to all 360 candidate scenes. It models *response topology*—an ordered photographic route in A Walk in the Park or a circular scene continuum in Scene Wheels—then tests whether fixed descriptors or frozen representations improve prediction of the chosen scene (Figure 1).

This question follows the target-confusability competition (TCC) model, which explains memory reports through target–candidate similarity (Schurgin et al., 2020). Deep features predict naturalistic Scene Wheel errors (Bates et al., 2024), and a priori human and model geometries predict target-specific working-memory distributions (Brady & Robinson, 2026). PRISM-Eval compares cognitive models and computer-vision representations on the same candidates, response-layout information, and held-out choices: does a learned representation predict *which* scene is confused with the target after simpler visual explanations are included?

Mixture models describe memory reports through a target-centered distribution and a uniform component (Zhang & Luck, 2008). PRISM-Channel adds visual similarity to explain *which* candidate is selected beyond this baseline. Here equal distance means equal displacement in the dataset's response coordinates, not equal perceptual similarity. Controlling these coordinates allows us to test whether images explain additional response structure; it does not assume that route positions or wheel angles form a perceptually uniform scale.

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

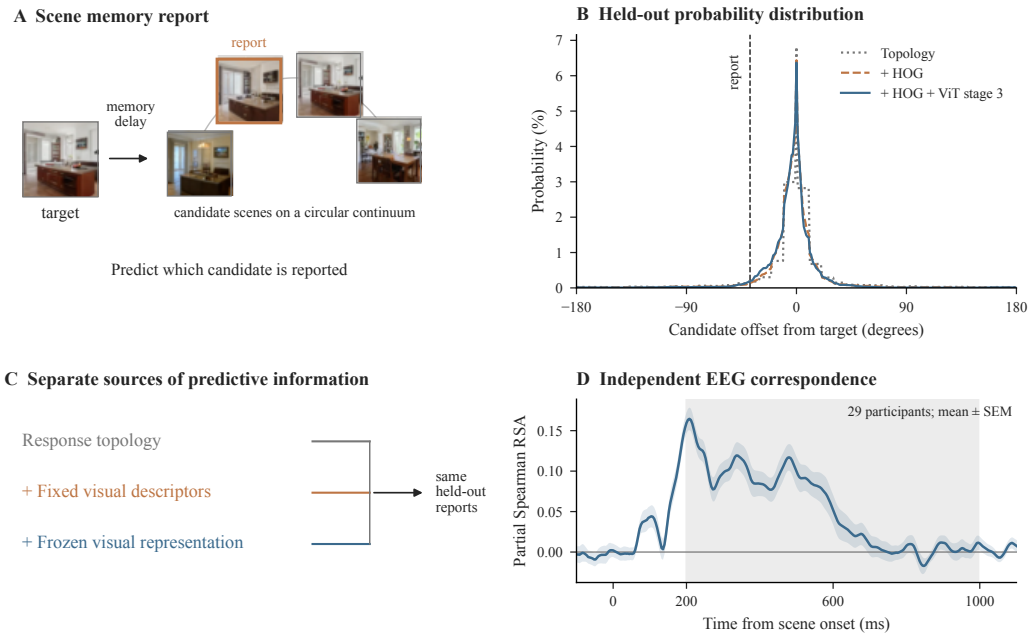

Figure 1: **From visual similarity to memory reports. A:** A participant maintains a target scene during a memory delay and selects one candidate; example stimuli are from Scene Wheels (Bates et al., 2024). **B:** Held-out 360-way predictions for the first trial by original index in the wheel-1 test set, from topology, topology plus HOG, and topology plus HOG and Vision Transformer (ViT) stage 3. The example was selected a priori and all parameters were fixed on development responses (Appendices B.3–B.4). **C:** Matched comparisons separate response topology, fixed image descriptors, and representations frozen prior to test evaluation. **D:** Empirical stage-3 partial RSA with independent EEG after HOG and RGB removal (29 participants; shading: SEM).

Which visual similarities should the memory model use? CCGD learns three weights over offset-centered frozen similarities through regularized response likelihood. A fitted equal-weight ensemble separates relative-weight learning from ensembling. The weights describe conditional predictive contributions, not neural activity or direct measurements of cognitive processes.

We test prediction on held-out routes and wheels, termed *environments*, and assess neural convergence using independent electroencephalography (EEG).

The paper makes four contributions:

1. **Predicting complete memory-report distributions.** PRISM-Eval compares cognitive and visual models on report identity in new participants, targets, and environments.
2. **Strong controls isolate visual geometry.** Fixed descriptors explain most pooled-encoder alignment, while intermediate ViT geometry adds behavioral information and predicts independent EEG structure.
3. **A visual channel models error identity.** PRISM-Channel adds target-specific visual evidence to a memory baseline, with an optional constraint preserving central probabilities and a target-specificity control.
4. **Learning which visual similarities matter.** A convex three-weight model improves prediction in new environments across two behavioral datasets and separates the benefit of combining visual models from the additional benefit of learning their relative weights.

2 Related Work

Similarity and memory errors. The target-confusability competition (TCC) model scales target–candidate similarity by memory strength, adds noise, and selects the strongest match (Schurgin et al.,

2020). Channel instead exponentially reweights a baseline distribution, optionally protects its center, and mixes back with it. These are different probability constructions, not a storage–decision dichotomy: TCC itself includes a decision rule. Ideal-observer, variable-precision, and population-coding accounts instead emphasize uncertainty, resource allocation, and noise (Sims et al., 2012; van den Berg et al., 2012; Bays, 2014; Bays et al., 2024). Perceptual similarity alone does not always recover working-memory error functions (Tomić & Bays, 2024), but richer geometries from human judgments and pretrained networks can predict complete, target-specific errors (Bates et al., 2024; Brady & Robinson, 2026). Competitive retrieval and interference theories also predict structured confusions (Oberauer & Lin, 2017). Because similarity can produce attraction, repulsion, or grouping depending on context (Lin & Luck, 2009; Chunharas et al., 2022; Czoschke et al., 2023), our outcome is the whole response distribution rather than a single error score. PRISM-Channel connects these traditions through an explicit mixture-model baseline and candidate-specific visual evidence. This uses the mixture model to describe baseline response spread without requiring its uniform component to identify a discrete psychological guessing state.

Comparing visual representations. Human–model comparisons depend on the dataset, objective, layer, and controls (Muttenthaler et al., 2023; Demircan et al., 2024); better computer vision need not mean better biological modeling (Linsley et al., 2026). Contextual perception distinguishes helpful from distracting context (Ying et al., 2023). Analogously, PRISM-Eval separates response layout, fixed visual structure, and learned geometry to locate the information underlying predictive improvements.

Learning visual geometry. The Zhang–Luck model mixes target-centered and uniform responses (Zhang & Luck, 2008); circular mixture regression lets angular components depend on covariates (Skhosana & Nakhaei Rad, 2026). Channel retains a mixture baseline but changes candidate odds through visual evidence. CCGD combines fixed similarities, as in multiple-kernel learning (Gönen & Alpaydın, 2011; Rakotomamonjy et al., 2008), using conditional multinomial likelihood (McFadden, 1973). Its cognitive log-probability offset, offset-centered visual scores, and compact learned readout specialize established statistical tools to memory reports. Unlike nonnegative kernel mixtures, its signed conditional-logit coefficients need not sum to one; it is not a new general-purpose optimizer.

Visual models of behavior and brain. Human–model comparisons should preserve the relevant behavioral observable (Garimella et al., 2024; Garcia et al., 2025). RTify models decision timing beyond mean accuracy (Cheng et al., 2024); PRISM-Eval instead predicts memory-report identity beyond mean error and evaluates pretrained visual networks alongside RGB histograms, image pixels, HOG and Gabor descriptors, and randomly initialized networks under a common likelihood.

Behavioral prediction and neural mechanism. Training recurrent networks on complete working-memory errors can generate hypotheses about error-producing dynamics (Radmard et al., 2026). The population-vector model represents scenes as noisy neural vectors (Kiat & Luck, 2026); PRISM instead evaluates conditional report distributions. Neural-alignment methods change a visual network using neural targets (Lu et al., 2026). PRISM tests whether visual similarities useful for predicting memory also explain independent EEG measurements, without fitting the representation to those measurements. Representational-similarity analysis (RSA) correlates pairwise model and EEG dissimilarities, allowing us to ask whether the geometry that predicts memory errors also appears in delay-period population activity.

3 METHOD

3.1 RESPONSE DISTRIBUTIONS

Trial i provides target t_i, report y_i, and 360 candidate scenes c. Let $\Delta(t_i, c)$ be signed response-display distance and $s_\phi(t_i, c)$ the cosine similarity from frozen encoder ϕ. For distance bin $b(\Delta)$, let h_b be the development-estimated smoothed log response rate (Appendix B.2). The topology model fits its strength:

$$\eta_i^D(c) = \beta_D h_{b(\Delta(t_i,c))}. \tag{1}$$

The visual model adds one standardized similarity term:

$$\eta_i^{DV}(c) = \beta_D h_{b(\Delta(t_i,c))} + \beta \widetilde{s}_\phi(t_i, c). \tag{2}$$

The fitted weights $\beta_D, \beta \geq 0$ multiply the fixed distance profile and $\widetilde{s}_\phi = (s_\phi - \mu_{\text{dev}})/\sigma_{\text{dev}}$, standardized over development trial–candidate pairs. Scores become probabilities through a softmax with uniform-guessing rate $0 \leq \lambda \leq .30$:

$$p(y_i = c \mid t_i) = (1 - \lambda)\frac{\exp(\eta_i(c))}{\sum_d \exp(\eta_i(d))} + \frac{\lambda}{360}. \tag{3}$$

Both models include response distance. Their likelihood contrast tests whether image similarity adds predictive information beyond that shared description.

3.2 Representation Evaluation

A representation supplies an embedding for every candidate image, without behavioral outcomes. PRISM-Eval fits topology-only and topology-plus-geometry models on development environments, then compares their negative log-likelihood (NLL) on held-out 360-way reports. This common output permits comparisons across architectures, layers, and training objectives. Appendix B.4 specifies candidate-order and feature-alignment requirements.

3.3 PRISM-Channel

The *cognitive baseline* predicts response probabilities from memory precision and uniform guessing before visual similarities are introduced. For participant s, radius r, and target $t = t_i$, it follows the Zhang–Luck mixture model (Zhang & Luck, 2008):

$$q_{s,r}(c \mid t) = (1 - g_{s,r})\,\mathrm{VM}_{360}(c - t; \kappa_{s,r}) + g_{s,r}/360. \tag{4}$$

The normalized discrete von Mises component has concentration $\kappa_{s,r} \geq 0$; $g_{s,r} \in [0, 1)$ is the uniform-guessing weight. Channel asks how visual evidence changes candidate odds beyond this baseline. For each frozen encoder, subtract the mean cosine similarity at each signed response offset, then divide by the residual SD across the wheel–radius environment. The mean of the ViT, ResNet, and ConvNeXt stage-3 residuals is $z(t, c)$. The visual channel proposes

$$v_s(c \mid t) \propto q_{s,r}(c \mid t) \exp\{T_s z(t, c)\}, \qquad p_s = (1 - a_s)q_{s,r} + a_s \mathcal{S}_{.50}(q_{s,r}, v_s). \tag{5}$$

$T_s > 0$ sets the strength of visual reweighting, and $a_s \in [0, 1]$ is its fitted mixing weight: $a_s = 0$ returns the memory baseline, whereas $a_s = 1$ uses the visual proposal after applying $\mathcal{S}_{.50}$. This optional P50 constraint preserves every probability in the smallest target-centered region containing at least half the baseline mass, and preserves total mass outside it. Each trial is evaluated using parameters fitted to the opposite balanced trial half. Comparisons share mixing weights across observers, remove P50, or permute target identities while preserving the similarity distribution at every signed offset. This target-specificity control tests correctly matched evidence, not a unique cognitive mechanism. P50 preserves each central candidate probability and total outer mass, not the complete error-magnitude distribution. Appendix D gives the definitions and fitting procedures. Candidates with the same baseline probability can receive different visual evidence, without treating response distance as perceptual similarity.

3.4 Cognitive-Channel Geometry Distillation

CCGD learns visual evidence for the memory model, compressing behavioral confusions into three image-computable similarity weights. Let $\mathbf{z}_i(c) \in \mathbb{R}^3$ contain the three residual similarities at t_i for trial i, and let $q_i(c) > 0$ be a cognitive baseline fitted on development responses. Define

$$\begin{aligned} p_{\mathbf{w}}(c \mid i) &= \frac{q_i(c) \exp\{\mathbf{w}^\top \mathbf{z}_i(c)\}}{\sum_d q_i(d) \exp\{\mathbf{w}^\top \mathbf{z}_i(d)\}}, \\ \widehat{\mathbf{w}} &= \arg\min_{\mathbf{w}} \left[-\frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \log p_{\mathbf{w}}(y_i \mid i) + \frac{\rho}{2} \|\mathbf{w}\|_2^2 \right]. \end{aligned} \tag{6}$$

Here y_i is the observed report, $\mathcal{T}$ contains training trials, and $\rho > 0$. The Hessian is a mean feature covariance under $p_{\mathbf{w}}$ plus ρI_3 (I_3 is the identity), giving a unique optimum. We fit three signed coefficients with deterministic L-BFGS; image features stay fixed. Zero weights recover q exactly. The evaluated CCGD uses the unrestricted tilt in Equation 6, which can change error spread as well as identity. Combining its learned geometry with P50 would be a separate test.

To distinguish learning from ensembling, a constrained control fixes the relative weights equal and fits only their common strength, using $z_{\mathrm{eq}} = (z_1 + z_2 + z_3)/\sqrt{3}$ and the same penalty family. This unit-norm direction matches penalties at equal coefficients for any given ρ. The comparison therefore asks whether learning *which visual geometry to emphasize* adds to the benefit of using several geometries.

4 Experiments

4.1 Behavioral Datasets

A Walk in the Park contains 15,035 trials from 114 participants, 1,080 targets, and three photographed routes (Westebbe et al., 2024). Each route is an ordered sequence of outdoor photographs from one walk. We hold out each complete route in turn; additional folds hold out targets or participants. **Scene Wheels** contains 5,400 trials from 20 participants, five circular continua of generated scenes, and five radii (Bates et al., 2024). We hold out each wheel and also evaluate participant-held-out folds. Each trial is scored once per evaluation protocol. Channel fits participant–radius precision and guessing, with optional observer-specific visual weights; CCGD fits shared visual weights. Participant resampling captures individual variability, not individual neural or psychological traits.

4.2 Representations and Controls

The declared representation bank contains eight pretrained encoders: ResNet50, ConvNeXt-B, ViT-B/16, Swin-B, DINOv2-B/14, MAE-B/16, CLIP-B/16, and SigLIP-B/16 (He et al., 2016; Liu et al., 2022; Dosovitskiy et al., 2021; Liu et al., 2021; Oquab et al., 2024; He et al., 2022; Radford et al., 2021; Zhai et al., 2023). Controls include red–green–blue (RGB) color histograms, downsampled pixels, randomly initialized networks, histograms of oriented gradients (HOG) (Dalal & Triggs, 2005), and Gabor features. The strongest fixed family contains six HOG/Gabor variants fitted together. We sample four predefined depth stages from ResNet50, ConvNeXt-B, and ViT-B/16; ViT stage 3 is the late-intermediate output after transformer block 9, before the final embedding.

For each held-out route or wheel, encoder, layer, and ridge are selected through five target-disjoint folds formed only from the remaining environments. The selected configuration is refit on all development trials before test scoring, so test responses never enter selection (Figure 3A; Appendix B). CCGD similarly fits environment-excluded cognitive baselines and visual weights to predict the omitted route or wheel, including its unseen Scene Wheels participants, whereas Channel cross-fits each observer across balanced trial halves. PRISM-Eval selects $\xi \in \{.001, .01, .1\}$ within inner target folds. Primary CCGD uses nested selection of $\rho \in \{.01, .1, 1\}$ within outer-training environments, while the fixed setting ($\rho = 1$) is evaluated separately as a historical exploratory benchmark (Appendices D, F, and G).

4.3 Evaluation

PRISM-Eval, Channel, and CCGD test, respectively, visual increments beyond topology and fixed descriptors, prediction for calibrated observers, and prediction in excluded environments using environment-excluded baselines. EEG tests correspondence in independent participants and scenes without EEG-fitted visual coefficients; Appendix D.5 compares the behavioral components under a common task and calibration budget.

The primary statistic is paired held-out NLL gain, $\mathrm{NLL}(D) - \mathrm{NLL}(DV)$. PRISM-Eval gives each trial equal weight; crossed participant–target bootstrap intervals account for repeated participants and images. Secondary analyses test calibration, response rank, participant-level consistency, error-only trials, and distance controls. Cognitive comparators—Zhang–Luck (Zhang & Luck, 2008), fixed and fitted DNN–TCC, and the noisy-winner TCC implementation of Bates et al. (2024)—all

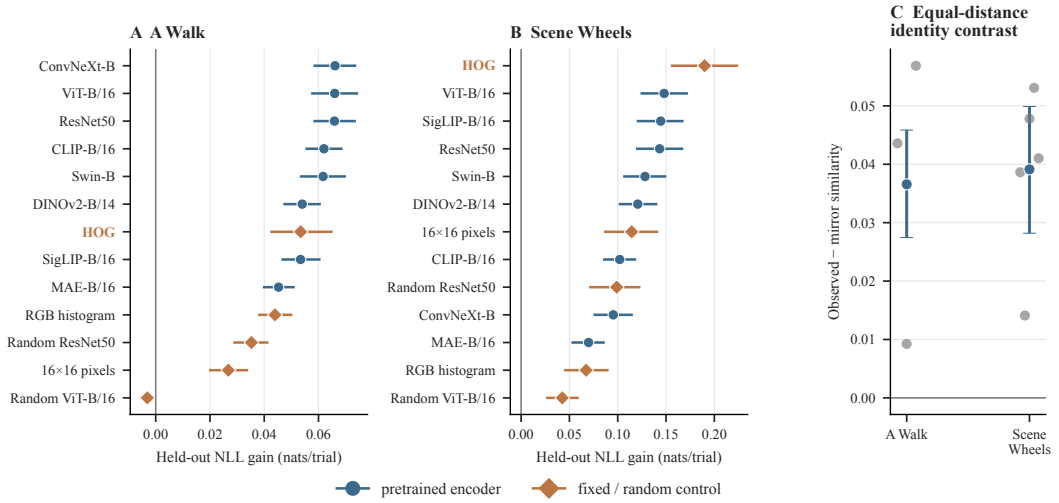

Figure 2: **A–B:** Held-out gain over response topology for pooled neural representations and fixed controls. HOG is competitive on A Walk and strongest on Scene Wheels. **C:** Selected scenes are more similar to their targets than alternatives at the same distance on the opposite side. Intervals are 95% intervals; dots are routes or wheels. Equal distance refers to response coordinates, not perceptual equivalence.

predict the same 360-way distributions. Channel and CCGD use equal-participant mean gains, participant bootstrap intervals, and sign-flip tests; Channel's three model contrasts are Holm-corrected. CCGD intervals condition on the available environments and fitted procedures. Appendices B.5, C, D, F, and G report metric definitions, correction families, and resampling budgets.

4.4 EEG Evaluation

The independent Kiat–Luck archive provides EEG similarities for 29 participants and 50 natural scenes (Kiat & Luck, 2026). We compare the behaviorally tested ViT stage-3 representation with EEG dissimilarity throughout the 200–996-ms interval after scene onset (0–796 ms after offset). Participant-specific Spearman RSA removes HOG and RGB dissimilarity jointly; sign-flip tests with Holm correction cover six neural comparisons. Appendix E specifies data preparation, the partial-RSA procedure, the time course, and participant-level inference.

5 Results

5.1 Pooled Representations

Against response topology alone, all eight pooled encoders improve held-out NLL on both datasets. A Walk gains range from 0.0452 to 0.0661 nats/trial; Scene Wheels gains range from 0.0680 to 0.1480. Every route, wheel, and participant-held-out mean is positive. The fixed controls change the interpretation (Figure 2): HOG gains 0.05344 nats/trial on A Walk and 0.18993 on Scene Wheels, where it exceeds every pooled learned representation (best: ViT-B/16, 0.14803). Only ConvNeXt-B, ResNet50, and ViT-B/16 exceed every control on A Walk, and none does so on Scene Wheels. Much of the pooled signal is therefore local edge-orientation structure, not a property unique to pretrained visual features.

5.2 Intermediate Visual Features

Fitting the six fixed image descriptors together gives the learned representation a demanding baseline (Figure 3). Adding the fixed ViT-B/16 stage-3 similarity improves trial-weighted held-out NLL by 0.02897 nats/trial on A Walk (crossed 98.75% CI [0.01928, 0.03875]) and 0.01880 on Scene Wheels ([0.00224, 0.03506]). These likelihood gains mean that the full model assigns, on geometric average,

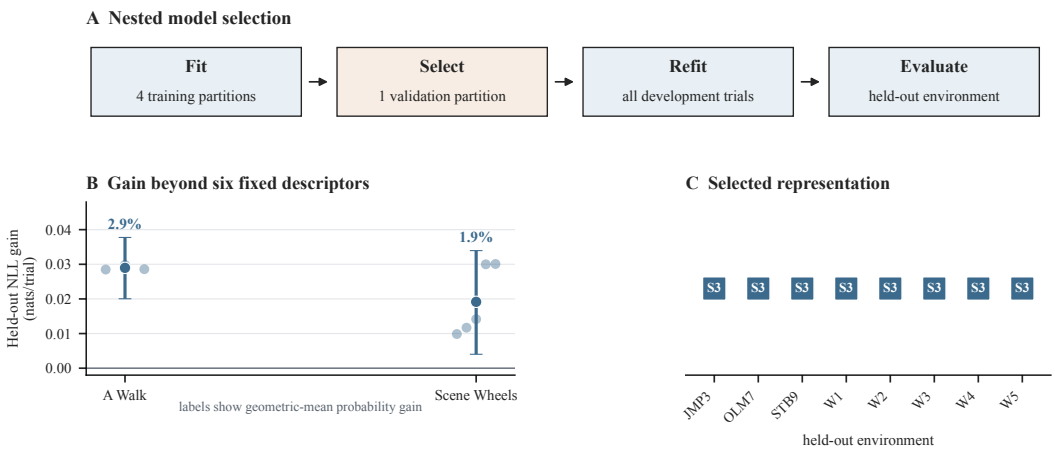

Figure 3: **A:** Five target-disjoint folds within the training environments select the encoder, layer, and regularization strength before evaluation on the held-out route or wheel. **B:** Mean held-out NLL gain from adding the cross-validation-selected neural representation to topology and six jointly fitted HOG/Gabor descriptors. Trials receive equal weight; bars are 97.5% per-dataset bootstrap intervals and dots are held-out routes or wheels. **C:** The representation selected for each of the eight held-out environment tests. Appendix B gives the full selection and resampling protocol.

2.94% and 1.90% more probability to the response a participant actually made. All three routes and all five wheels improve.

Of the NLL improvement from topology to the complete fixed-stage model, the six descriptors account for 67.1% on A Walk and 91.7% on Scene Wheels; intermediate stage-3 geometry contributes the remaining 32.9% and 8.3%. Its absolute gains are supported by the simultaneous confidence intervals above; the percentages are descriptive shares within this model sequence, not fractions of all predictable behavior.

Repeating representation selection separately for every test environment yields the same layer (Figure 3C). The development partitions choose ViT stage 3 in all eight tests. Adding the selected representation gives a trial-weighted gain of 0.02897 nats/trial on A Walk (97.5% per-dataset CI [0.02004, 0.03775]) and 0.01916 on Scene Wheels ([0.00402, 0.03395]), with a positive mean in every held-out route and wheel. The participant-weighted analysis is also positive in both datasets (106/114 A Walk participants and 16/20 Scene Wheels participants). Stage 3 outperforms the final ViT embedding in both datasets, and the effect remains positive after correct reports are removed and under all five alternative distance models. Appendices C.2–C.4 report participant-level consistency, error-only and distance-sensitivity results, and the equal-distance identity contrast. Gains also survive all seven lapse-cap, bin-width, and smoothing settings (Appendix K.1).

5.3 Cognitive Model Comparison

The distance + HOG + stage-3 model has the lowest mean held-out NLL in the cognitive-comparator family, which uses one HOG descriptor. Corrected simultaneous inference supports gains over Zhang–Luck and fixed and fitted DNN–TCC in both datasets, and over the quadrature-evaluated noisy-winner model in Scene Wheels (gain .16364; lower bound .01586). Quadrature preserves all eight parameter selections; increasing its resolution from 100 to 200 nodes changes mean NLL by less than 10^{-6} nats/trial. Appendices C.5–C.6 retain the complete corrected comparison family and numerical sensitivity.

5.4 PRISM-Channel

The primary PRISM-Channel test uses the circular Scene Wheels baseline. In cross-fitted evaluation, the variant with observer-specific weights and P50 gains 0.03757 nats/trial over that baseline (95% participant bootstrap [0.02040, 0.05743], one-sided $p = 1.0 \times 10^{-5}$; 18/20 observers), with

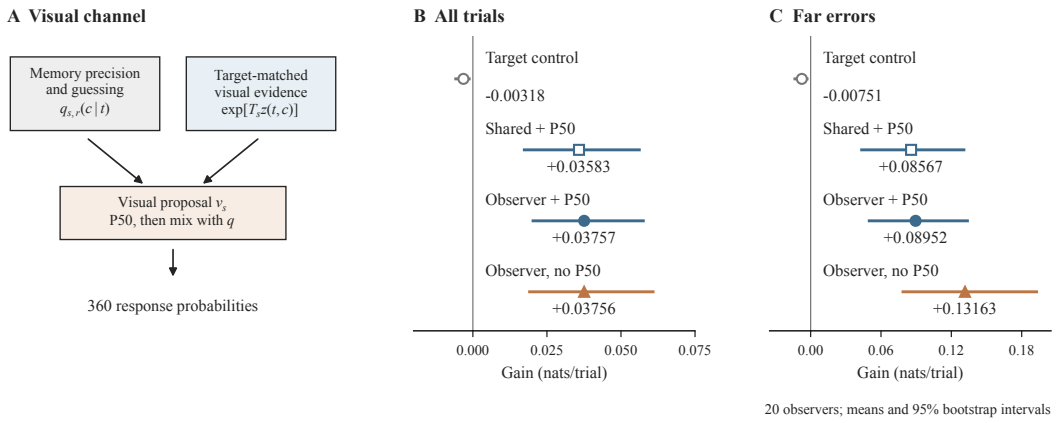

Figure 4: **Visual identity augments the memory baseline. A:** A memory baseline supplies precision and uniform guessing; visual residuals frozen prior to test evaluation redistribute probability among target-matched identities. **B–C:** Log-likelihood gain over the baseline on all held-out trials and on far errors outside its smallest target-centered 50%-mass region. Points and labels are participant means; bars are 95% participant-bootstrap intervals. The target-specificity control breaks target matching while preserving offset-wise score marginals; P50 preserves total far-error mass. All parameters are fitted on the opposite trial half (Appendix D.3).

every wheel positive. The trial-weighted, baseline-defined far-tail gain is 0.09260 (participant mean .08952; 95% CI [0.05003, 0.13396], $p = 1.4 \times 10^{-5}$; 18/20).

The target-specificity control changes target identity while preserving visual-score marginals at every signed offset. This control loses 0.00318 nats/trial; PRISM-Channel exceeds it by 0.04076 ([0.02376, 0.06037], Holm $p = 1.2 \times 10^{-5}$; 19/20 observers). Thus frozen visual geometry helps because the evidence is matched to the remembered target, not because it merely reproduces the response circle. Shared mixing weights also support positive prediction. Removing P50 preserves overall gain and increases the far-error estimate (Figure 4). These comparisons support target-matched visual evidence within the memory model; they do not establish observer-specific weights or the P50 constraint as necessary.

Channel gains survive optimized and flexible memory baselines (.04118 and .04392 nats/trial), exact-distance conditioning (.02395), and 20 fixed target-specificity perturbations (Appendix D.4, including intervals). These checks isolate visual information beyond baseline spread. On the matched held-out-wheel task, shared Channel has the lowest mean NLL among the tested models, without observer calibration (Appendix D.5). Gains persist across all four protection levels and both tuning regimes (16 contrasts, Holm $p \leq .00288$). Linear A Walk Channel also improves prediction: +.01680 nats [.01048,.02252] (Appendices K.1, K.3).

5.5 Learning Visual Similarity Weights

With nested penalty selection, CCGD improves trial-weighted held-out log likelihood over its cognitive baseline by 0.01956 nats/trial on A Walk and 0.02416 on Scene Wheels (Appendix G.7), using Section 3.4's cognitive baselines, not the six-descriptor PRISM-Eval model. The primary unequal-versus-equal comparison gives participant-mean gains of .000516 (pointwise 95% CI [−.001607,.003016]) and −.002267 ([−.004991,.000093]), respectively. Both intervals include zero; nested selection does not establish an advantage for learned relative weights (Appendix G.6).

The fixed-$\rho = 1$ benchmark gives positive learning increments of .000785 and .000391 nats/trial: .079% and .039% increases in geometric-mean observed-response probability, using $100(\exp(\Delta) - 1)$.

Matched readouts add the three neural features beyond six jointly fitted descriptors: participant-mean gains are .01189 [.00943,.01417] and .00474 [.00141,.00839], both surviving six-test Holm

correction (Appendix K.2). Flexible-baseline and exact-distance gains remain positive (Appendices G.7–G.8).

In an independent two-choice memory dataset (Pettini et al., 2026), CCGD's frozen weight direction transfers across held-out participants and scene families (240 and 108 total). Its gain over equal weighting is .000600 nats (crossed 95% CI [.000140,.001083]; Holm $p = 1.5 \times 10^{-5}$, simulation-limited). This is adapted recognition transfer, not unchanged 360-way prediction (Appendix K.4).

5.6 Neural Convergence

Mean timewise stage-3 RSA after jointly removing HOG and RGB dissimilarity is 0.05394 (25/29 participants; Holm $p = 6.0 \times 10^{-5}$). Differences in raw timewise RSA, averaged within participant over the same window, favor stage 3 over HOG by .02638 and RGB by .01589; both survive correction. This supports shared delay-period structure beyond edge orientation and color in independent participants and scenes, not a causal neural mechanism.

For CCGD, averaging weights fitted to the two behavioral datasets gives a visual geometry without EEG-supervised coefficients. Its partial RSA on the delay-averaged EEG matrix is .06027 versus .05511 for ViT stage 3 (paired gain .00516, 95% interval [.00027,.00993], nominal $p = .02489$; 20/29 participants). Both learned and equal-weight combinations capture cross-modal structure (Appendix G.5). Controlling all six descriptors retains partial RSA of .05374 for stage 3, .05878 for CCGD and .06010 for equal weighting (scene-label permutation Holm $p = 3.0 \times 10^{-4}$). CORnet-S IT features also show positive descriptive alignment (.04447); Appendix K.2 gives paired uncertainty, without establishing model superiority.

6 Discussion

Scene-memory reports reflect response structure and visual similarity. HOG explains much pooled-encoder alignment (Dalal & Triggs, 2005; Geirhos et al., 2019), while intermediate ViT adds information beyond six fixed controls and the final embedding; a post-development EfficientNet-B0 portability test is also positive in all routes and wheels (Appendix I). Stage 3 may retain scene-specific spatial detail or better match ventral computations (Nonaka et al., 2021), but alignment does not distinguish these possibilities (Conwell et al., 2024).

PRISM-Channel joins accounts of response spread and candidate similarity (Sims et al., 2012; Schurgin et al., 2020; Bays et al., 2024) in one response distribution. Target permutation preserves offset-wise visual score distributions but removes the predictive benefit: matching evidence to the target matters beyond response topology. Shared weights suffice, and P50 is an optional restriction on central and tail probabilities, not a necessary condition for positive prediction. Neither better fit nor these equations localize storage versus readout. Crossed delay–response-cost simulations recover restricted rivals in 95–100% of signal-bearing cases (Appendix K.5), motivating tests of human mechanisms.

CCGD fits three coefficients over frozen encoders under a cognitive log-probability offset. Convex nesting makes it tractable, not a general-purpose optimizer. Ensembling supplies most of the gain; relative-weight learning adds a smaller increment under the fixed exploratory benchmark, with positive adapted external-memory transfer. The regularized coefficients describe conditional changes in candidate odds, not brain activity. Channel and CCGD require neither encoder fine-tuning nor neural measurements from the observer.

Natural scenes challenge theories built around discrete features (Luck & Kiat, 2024). PRISM connects image-computable memory models (Bates et al., 2024) to report identity and neural geometry. The cognitive constraints are that intermediate visual information survives tested descriptors, target matching predicts identity beyond response distance, and CCGD's learned weights support adapted cross-task transfer. These findings can guide scene sets matched in response distance but chosen to elicit divergent model predictions (Golan et al., 2020). The EEG correspondence and adapted recognition transfer extend this framework without identifying cortical sources or causal stages. PRISM tests scene-memory theories and evaluates vision representations through human memory confusions.

Reproducibility and Ethics Statement

We reanalyze existing behavioral and EEG data and collect no new human data. The supplement records data sources, preprocessing, folds, statistical tests, and complete tables. Scene Wheels images are reproduced with CC BY 4.0 attribution. Appendix H records data access and licensing.

AI Use Statement

The authors are responsible for the research design, algorithm development, implementation, theoretical-model development, analysis, literature search, interpretation, and all claims in this submission. Under author direction, generative AI tools assisted analysis, literature search, figure preparation, and manuscript editing. All reported participant observations come from the cited behavioral and EEG archives; AI tools did not collect participant data. Reproducible code and numerical audits accompany the analyses.

References

Christopher J. Bates, George A. Alvarez, and Samuel J. Gershman. Scaling models of visual working memory to natural images. *Communications Psychology*, 2:3, 2024. doi: 10.1038/s44271-023-00048-3.

Paul M. Bays. Noise in neural populations accounts for errors in working memory. *Journal of Neuroscience*, 34(10):3632–3645, 2014. doi: 10.1523/JNEUROSCI.3204-13.2014.

Paul M. Bays, Sebastian Schneegans, Wei Ji Ma, and Timothy F. Brady. Representation and computation in visual working memory. *Nature Human Behaviour*, 8(6):1016–1034, 2024. doi: 10.1038/s41562-024-01871-2.

Timothy F. Brady and Maria M. Robinson. Predicting visual working memory errors with a priori measures of representational geometry from human and artificial systems. PsyArXiv preprint, 2026. URL `https://osf.io/preprints/psyarxiv/k3xu6_v1`.

Timothy F. Brady, Talia Konkle, and George A. Alvarez. Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, 138(4):487–502, 2009. doi: 10.1037/a0016797.

Yu-Ang Cheng, Ivan Felipe Rodriguez, Sixuan Chen, Kohitij Kar, Takeo Watanabe, and Thomas Serre. RTify: Aligning deep neural networks with human behavioral decisions. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130485–130510, 2024.

Chaipat Chunharas, Rosanne L. Rademaker, Timothy F. Brady, and John T. Serences. An adaptive perspective on visual working memory distortions. *Journal of Experimental Psychology: General*, 151(10):2300–2323, 2022. doi: 10.1037/xge0001191.

Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15:9383, 2024. doi: 10.1038/s41467-024-53147-y.

Stefan Czoschke, Benjamin Peters, Jochen Kaiser, and Christoph Bledowski. Interference between items stored for distinct tasks in visual working memory. *Attention, Perception, & Psychophysics*, 85:1461–1473, 2023. doi: 10.3758/s13414-023-02657-w.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pp. 886–893, 2005. doi: 10.1109/CVPR.2005.177.

Can Demircan, Tankred Saanum, Leonardo Pettini, Marcel Binz, Blazej M. Baczkowski, Christian F. Doeller, Mona M. Garvert, and Eric Schulz. Evaluating alignment between humans and neural network representations in image-based learning tasks. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3890.

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

Kathy Garcia, Emalie McMahon, Colin Conwell, Michael F. Bonner, and Leyla Isik. Modeling dynamic social vision highlights gaps between deep learning and humans. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b1bdb0f22c9748203c62f29aa297ac57-Abstract-Conference.html.

Manjulata Garimella, Denizhan Pak, Justin N. Wood, and Samantha M. W. Wood. A newborn embodied turing test for comparing object segmentation across animals and machines. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/01d78b294d80491fecddea897cf03642-Abstract-Conference.html.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Tal Golan, Prashant C. Raju, and Nikolaus Kriegeskorte. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, 117(47): 29330–29337, 2020. doi: 10.1073/pnas.1912334117.

Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011. URL https://www.jmlr.org/papers/v12/gonen11a.html.

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

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

John E. Kiat and Steven J. Luck. A population vector model of visual working memory for real-world scenes. *Journal of Experimental Psychology: General*, 155(5):1257–1281, 2026. doi: 10.1037/xge0001921.

Po-Han Lin and Steven J. Luck. The influence of similarity on visual working memory representations. *Visual Cognition*, 17(3):356–372, 2009. doi: 10.1080/13506280701766313.

Drew Linsley, Pinyuan Feng, and Thomas Serre. Better artificial intelligence does not mean better models of biology. *Trends in Cognitive Sciences*, 30(7):599–610, 2026. doi: 10.1016/j.tics.2025.11.016.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Zitong Lu, Yile Wang, and Julie D. Golomb. Achieving more human brain-like vision via human EEG representational alignment. *Communications Biology*, 9:463, 2026. doi: 10.1038/s42003-026-09685-w.

Steven J. Luck and John E. Kiat. Visual working memory for natural scenes: Challenges and opportunities. *Cognitive Processing*, 25(Suppl 1):73–78, 2024. doi: 10.1007/s10339-024-01213-0.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. Technical report, University of California, Berkeley, 1973. URL https://escholarship.org/uc/item/0p99072q.

Lukas Muttenthaler, Jonas Dippel, Lorenz Johannes Linhardt, Robert A. Vandermeulen, and Simon Kornblith. Human alignment of neural network representations. In *International Conference on Learning Representations*, 2023.

Soma Nonaka, Kei Majima, Shuntaro C. Aoki, and Yukiyasu Kamitani. Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *iScience*, 24(9):103013, 2021. doi: 10.1016/j.isci.2021.103013.

Klaus Oberauer and Hsuan-Yu Lin. An interference model of visual working memory. *Psychological Review*, 124(1):21–59, 2017. doi: 10.1037/rev0000044.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Leonardo Pettini, Carsten Bogler, Christian Doeller, and John-Dylan Haynes. Synthesis and perceptual scaling of high-resolution naturalistic images using Stable Diffusion. *Behavior Research Methods*, 58:24, 2026. doi: 10.3758/s13428-025-02889-8.

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

Puria Radmard, Paul Bays, and Máté Lengyel. Setting up for failure: Automatic discovery of the neural mechanisms of cognitive errors. In *International Conference on Learning Representations*, 2026.

Alain Rakotomamonjy, Francis R. Bach, Stéphane Canu, and Yves Grandvalet. Simplemkl. *Journal of Machine Learning Research*, 9:2491–2521, 2008. URL `https://www.jmlr.org/papers/v9/rakotomamonjy08a.html`.

Mark W. Schurgin, John T. Wixted, and Timothy F. Brady. Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour*, 4(11):1156–1172, 2020. doi: 10.1038/s41562-020-00938-0.

Chris R. Sims, Robert A. Jacobs, and David C. Knill. An ideal observer analysis of visual working memory. *Psychological Review*, 119(4):807–830, 2012. doi: 10.1037/a0029856.

Sphiwe B. Skhosana and Najmeh Nakhaei Rad. Model-based clustering using a new mixture of circular regressions. *Advances in Data Analysis and Classification*, 2026. doi: 10.1007/s11634-026-00673-w.

Robert Taylor and Paul M. Bays. Efficient coding in visual working memory accounts for stimulus-specific variations in recall. *Journal of Neuroscience*, 38(32):7132–7142, 2018. doi: 10.1523/JNEUROSCI.1018-18.2018.

Ivan Tomić and Paul M. Bays. Perceptual similarity judgments do not predict the distribution of errors in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 50(4):535–549, 2024. doi: 10.1037/xlm0001172.

Ronald van den Berg, Hongsup Shin, Wen-Chuang Chou, Ryan George, and Wei Ji Ma. Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, 109(22):8780–8785, 2012. doi: 10.1073/pnas.1117465109.

Leo Westebbe, Yibiao Liang, and Erik Blaser. The accuracy and precision of memory for natural scenes: A walk in the park. *Open Mind*, 8:131–147, 2024. doi: 10.1162/opmi_a_00122.

Zhuofan Ying, Peter Hase, and Mohit Bansal. Adaptive contextual perception: How to generalize to new backgrounds and ambiguous objects. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1951.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Weiwei Zhang and Steven J. Luck. Discrete fixed-resolution representations in visual working memory. *Nature*, 453:233–235, 2008. doi: 10.1038/nature06860.

PRISM: PREDICTING COMPLETE SCENE-MEMORY REPORT DISTRIBUTIONS FROM VISUAL GEOMETRY

SUPPLEMENTARY MATERIAL

Anonymous authors
Paper under double-blind review

A DATA AND EVALUATION SPLITS

A.1 A WALK IN THE PARK

The processed A Walk dataset contains 15,035 trials from 114 participants, 1,080 target positions, and three photographed routes. Each route contains 360 legal candidates ordered along a line. We use linear signed offsets, with exact target choices assigned a dedicated distance bin. Leave-one-route-out evaluation holds out each complete route once. Five additional participant folds are group-disjoint, and five target folds keep all trials for a target in one fold.

A.2 SCENE WHEELS

Scene Wheels contains 5,400 trials from 20 participants, five circular scene continua, and five radii. Each wheel–radius ring contains 360 candidates, for $5 \times 5 \times 360 = 9,000$ unique 256×256 RGB images. We use circular shortest offsets with period 360, with dedicated target and 180-degree antipode bins. Three responses recorded as 360 are mapped to 0 before indexing. Leave-one-wheel-out evaluation holds out each wheel once; five participant folds are stratified by radius.

A.3 GENERALIZATION UNIT AND FOLD INTEGRITY

A route or wheel is an environment. Holding out an environment changes the image set, response topology, and participant composition available during fitting. We use *outer fold* only as shorthand for one test in which a complete environment is withheld and all remaining environments form the development set. In each such test, model coefficients and response-derived normalizers are fitted using development trials only. Representations frozen prior to test evaluation may be used to compute response-free similarities, but no response from the held-out environment enters coefficient fitting. For the declared PRISM-Eval bank, representation selection also excludes the held-out environment. Channel and CCGD use the calibration and fixed-penalty or separately nested protocols described in Appendices D, F, and G. Across each set of three route tests and five wheel tests, every trial is scored exactly once.

Each model predicts a normalized distribution over the same 360 candidates; the response log score is the log probability at the observed candidate. Participant, target, environment, and radius identifiers preserve the repeated-measures structure for paired comparisons and resampling.

Individual differences are handled differently by the three components. PRISM-Eval tests shared visual predictors and includes a participant-specific precision sensitivity analysis (Appendix C). Channel fits memory parameters for each participant and radius, then compares shared and observer-specific visual weights (Appendix D). CCGD estimates shared visual weights; its Scene Wheels baseline pools training observers by radius and predicts four new observers in each test wheel (Appendix G). These designs evaluate prediction and between-subject variability, not a model of stable individual traits.

B Representations and Complete-Distribution Models

B.1 Frozen Encoders and Fixed Descriptors

No image encoder is optimized using memory responses. Table 1 lists the eight pooled representations. The layer bank samples four depth stages from ResNet50, ConvNeXt-B, and ViT-B/16. ViT stage 3 uses the class-token (CLS) output after transformer block 9, before the final embedding.

Table 1: Frozen encoder bank.

Encoder	Training family	Pooled representation
ResNet50	Supervised convolutional	Global pooled
ConvNeXt-B	Supervised convolutional	Global pooled
ViT-B/16	Supervised transformer	CLS token
Swin-B	Supervised hierarchical transformer	Global pooled
DINOv2-B/14	Self-supervised transformer	CLS token
MAE-B/16	Masked autoencoding transformer	CLS token
CLIP-B/16	Image-text contrastive	Image embedding
SigLIP-B/16	Sigmoid image-text	Image embedding

Fixed controls include 32-bin histograms for each RGB channel (96 dimensions), bilinearly downsampled 16×16 RGB pixels centered by channel and divided by image-wide SD (768 dimensions), randomly initialized ResNet50 and ViT-B/16, and six edge or texture descriptors. The legacy HOG descriptor uses OpenCV's default settings on 64×128 grayscale inputs (3,780 dimensions). Three square HOG variants preserve the field of view by isotropically resizing to 64, 128, or 256 pixels and reflection-padding the shorter dimension. They use nine unsigned orientation bins, 2×2-cell blocks, one-cell stride, L2-Hys normalization, gamma correction, and cell widths of 8, 16, or 32 pixels, yielding 1,764 dimensions each. Multiscale HOG concatenates their separately normalized vectors (5,292 dimensions). The Gabor descriptor uses wavelengths 4, 8, 16, and 32 pixels; eight orientations; aspect ratio $\gamma = .5$; Gaussian-envelope width $\sigma = .56\lambda$ at wavelength λ; a 4×4 spatial grid; and mean absolute response plus RMS energy (1,024 dimensions).

B.2 Probability Model and Optimization

Trial i has target t_i and report y_i; candidate indices c, d range over $\{0, \ldots, 359\}$. A generic target is denoted t. Signed offset $\Delta(t, c)$ is $c - t$ on a route and the shortest circular offset on a wheel (the antipode has its own bin). For each development fold, let C_b count observed responses in bin b and E_b count all target-candidate exposures in that bin. Estimate

$$h_b = \log \frac{C_b + \alpha}{E_b + 2\alpha}, \qquad \alpha = 1. \tag{1}$$

These log rates are fixed during coefficient optimization, not independently optimized bin logits. The topology-only score is

$$\eta_i^D(c) = \beta_D h_{b(\Delta(t_i,c))}. \tag{2}$$

The topology-plus-visual score is

$$\eta_i^{DV}(c) = \beta_D h_{b(\Delta(t_i,c))} + \beta \widetilde{s}_\phi(t_i, c), \qquad \widetilde{s}_\phi(t_i, c) = \frac{s_\phi(t_i, c) - \mu_{\text{dev}}}{\sigma_{\text{dev}}}, \tag{3}$$

where s_ϕ is cosine similarity from encoder ϕ frozen prior to test evaluation; μ_{dev} and $\sigma_{\text{dev}} > 0$ are its mean and population SD over all development trial–candidate pairs. Invalid or near-constant scales ($\leq 10^{-12}$) are rejected. The bin width is 10 positions, with special target bins in both datasets and an additional antipode bin for Scene Wheels. The probability of candidate c is

$$p(y_i = c \mid t_i) = (1 - \lambda)\frac{\exp(\eta_i(c) - m_i)}{\sum_d \exp(\eta_i(d) - m_i)} + \frac{\lambda}{360}, \qquad m_i = \max_d \eta_i(d), \tag{4}$$

where subtracting the maximum is a numerically stable implementation of the same softmax and $0 \leq \lambda \leq .30$. All component coefficients are nonnegative. With $\mathbf{x}_i(c)$ containing h_b and the

standardized visual scores, and coefficient vector $\mathbf{b}$, fits minimize

$$-\frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}}\log p_{\mathbf{b},\lambda}(y_i \mid t_i) + \xi\|\mathbf{b}\|_2^2, \qquad \mathbf{b} \geq 0. \tag{5}$$

Here $\mathcal{T}$ is the development training set, $\eta_i(c) = \mathbf{b}^\top \mathbf{x}_i(c)$, and ξ is the ridge coefficient; the lapse rate λ is unpenalized. L-BFGS-B uses at most 80 iterations for the encoder census, pooled-control, single-descriptor-family, and earlier nested-unique analyses, and 150 for the joint six-descriptor and fully nested representation-selection analyses. The joint fixed-control model gives each of six fixed descriptors a fitted coefficient; adding a learned representation adds one more.

B.3 Nested Representation Selection

The candidate set consists of four stages from each of ConvNeXt-B, ResNet50, and ViT-B/16, crossed with ridge values $\{0.001, 0.01, 0.1\}$. For each held-out route or wheel, development targets are divided into five deterministic, target-disjoint partitions. Each partition serves as validation data once, while the other four fit the candidate. For every candidate pair, validation NLL is pooled over the five rotations with trial weighting. The minimum-NLL representation and ridge are selected via inner validation loops, refit on all development trials, and evaluated once on the held-out environment. Candidate order and then ridge order break exact ties. This procedure permits model selection within development data but keeps the held-out responses unavailable until final scoring. It evaluates a declared model family; it is not a new representation-learning algorithm. Main-paper Figure 3A summarizes this procedure; Figure 3B reports trial-weighted gains with 97.5% per-dataset bootstrap intervals (Bonferroni-adjusted for 95% simultaneous coverage), and Figure 3C records the selected representation in each outer test. All fitted parameters and selected hyperparameters are fixed before that test is scored.

B.4 Embedding Inputs and Example Predictions

Each encoder supplies one finite, nonzero feature vector per candidate image, in the same candidate order used by the behavioral data. Embedding inputs contain no responses. The evaluator computes cosine similarities and fits the shared response models on development trials. The resulting 360-way test distributions make representations comparable regardless of feature dimension or network architecture.

Main-paper Figure 1B displays the earliest original trial in the wheel-1 test set, selected a priori without reference to its model gain. Its target is at 60° on radius 16 and its observed report at 22° (an offset of -38°). The plotted probabilities come from the held-out topology, topology-plus-HOG, and topology-plus-HOG-plus-stage-3 models. Its coefficients and lapse rate were fixed on development responses prior to test-set scoring. This example uses one fixed descriptor, not the six-descriptor comparison in Appendix C. No parameters are refitted for the display, and no smoothing is applied to the 360 candidate probabilities. Aggregate comparisons, rather than this single trial, support the predictive claims. Matched comparisons in Figure 1C separate response topology, fixed image descriptors, and representations frozen prior to test evaluation.

B.5 Inference

The primary effect is the paired held-out NLL reduction $\mathrm{NLL}(D) - \mathrm{NLL}(DV)$, equivalently the mean improvement in observed-choice log probability. Intervals use a crossed participant–target pigeonhole bootstrap (Owen, 2007), preserving repeated observations from participants and images. Confidence levels distinguish each interval's nominal coverage from coverage across a contrast family. For nested selection, 97.5% intervals per dataset provide a Bonferroni family-wise target of 95%. They resample retained out-of-fold scores, not the historical model search. Participant-level consistency is assessed by first averaging trial effects within participant; one-sided sign-flip tests then weight participants equally. Holm correction is used within declared contrast families (Holm, 1979).

For test trials i, define the paired gain $d_i = \log p_A(y_i \mid t_i) - \log p_B(y_i \mid t_i)$, in nats. The trial-weighted gain is $|\mathcal{E}|^{-1}\sum_{i\in\mathcal{E}} d_i$. Participant weighting first computes $\bar{d}_s = |\mathcal{E}_s|^{-1}\sum_{i\in\mathcal{E}_s} d_i$ for participant s and then averages the $\bar{d}_s$ equally. Positive gain favors model A. Far-error analyses

restrict these sets to observed reports outside the opposite-half baseline's P50 region; they score the original probabilities, not a newly renormalized truncated model.

C Behavioral Evidence and Robustness

C.1 Pooled Representations and Fixed Controls

All pooled encoders improve on response topology in both datasets (Table 2). HOG changes the ranking: it is competitive in A Walk and exceeds every pooled learned representation in Scene Wheels (Table 3).

Table 2: Mean held-out NLL gain over response topology (nats/trial). The three columns use leave-one-route, leave-one-wheel, and participant-held-out folds.

Encoder	A Walk route	Scene wheel	Scene participant
ResNet50	.06585	.14335	.14379
ConvNeXt-B	.06606	.09542	.09917
ViT-B/16	.06595	.14801	.14897
Swin-B	.06168	.12822	.12915
DINOv2-B/14	.05401	.12075	.11982
MAE-B/16	.04525	.06805	.06740
CLIP-B/16	.06199	.10213	.10163
SigLIP-B/16	.05339	.14449	.14379

Table 3: Leave-one-environment-out NLL gain over topology for the strongest pooled encoder and all fixed or random controls.

Representation	A Walk gain	Scene Wheels gain
Best pooled learned (ConvNeXt/ViT)	.06609	.14803
HOG	.05344	.18993
16×16 pixels	.02677	.11445
RGB histogram	.04395	.06734
Random ResNet50	.03528	.09895
Random ViT-B/16	-.00308	.04263

C.2 Joint Controls and Nested Selection

Adding fixed ViT stage 3 to all six jointly fitted descriptors gives gains of .02897 nats/trial in A Walk (98.75% per-contrast CI [.01928, .03875]) and .01880 in Scene Wheels ([.00224, .03506]). The nested representation-selection pipeline independently selects ViT stage 3 for every held-out environment via inner development cross-validation (Table 4). The geometric-mean observed-choice probability gains are 2.94% and 1.90%. Relative to the full improvement over topology, fixed descriptors account for 67.1% in A Walk and 91.7% in Scene Wheels; stage 3 accounts for the remaining 32.9% and 8.3%. Stage 3 outperforming the final ViT embedding does not imply that every final ViT representation discards spatial information; the alignment result alone does not distinguish that explanation from other layer differences.

Table 4: Increment from the representation selected by development-set cross-validation beyond the six-descriptor model. Intervals have 97.5% coverage per dataset.

Dataset	Selected model	Mean gain	97.5% CI	Positive n
A Walk	3/3 ViT stage 3	.02897	[.02004, .03775]	106/114
Scene Wheels	5/5 ViT stage 3	.01916	[.00402, .03395]	16/20

Descriptive target-level summary. Using the same retained outer-test predictions, first average the selected-model minus six-descriptor log-probability gain within each target, then weight all observed targets equally. No model is refitted and no trials are excluded. The mean is .02977 nats/trial over 1,080 A Walk targets (746 positive, 334 negative) and .01957 over 4,295 Scene Wheels targets (2,327 positive, 1,968 negative). These descriptive means support a positive average across observed targets, not a new significance test or independent validation. In Scene Wheels, 3,310 targets have only one trial, so individual target effects are noisy.

C.3 Distance and Error-Only Sensitivity

Five alternatives replace the signed-bin topology while retaining the same folds, visual terms, and optimizer. Hierarchical variants add centered participant-specific precision; smooth variants use linear or circular kernels with the stated bandwidth. Bonferroni 99.5% intervals cover all ten dataset-by-distance tests (Table 5).

Table 5: Held-out stage-3 increment beyond HOG under stronger descriptions of response distance.

Distance model	A Walk Mean	99.5% CI	Scene Wheels Mean	99.5% CI
Binned + participant	.02295	[.01124, .03437]	.02721	[.01056, .04200]
Smooth, bandwidth 5	.02698	[.01690, .03709]	.02517	[.00777, .04068]
Smooth, bandwidth 10	.02520	[.01493, .03509]	.02718	[.00960, .04313]
Smooth, bandwidth 20	.05431	[.03904, .06829]	.02904	[.01169, .04543]
Smooth 10 + participant	.02509	[.01475, .03512]	.02772	[.01074, .04295]

Removing exact-target choices without refitting leaves 14,244 A Walk error trials and 5,030 Scene Wheels error trials. The full-minus-fixed-control gain is .03446 [.02658, .04229] in A Walk and .02570 [.01430, .03693] in Scene Wheels, with all three routes and all five wheels positive.

C.4 Equal-Distance Identity Contrast

For each non-target response with a legal equal-magnitude mirror, we compare target similarity to the selected scene with target similarity to that mirror. The aligned statistic averages standardized cosine differences over all eight encoders; the null averages 100 within-gallery image-label permutations (Table 6).

Table 6: Observed-minus-mirror target similarity at equal response distance.

Dataset	Included	Aligned mean	Permuted mean	95% CI of difference
A Walk	12,554	.03674	.00018	[.02746, .04586]
Scene Wheels	5,029	.03895	-.00018	[.02819, .04989]

C.5 Matched Cognitive Comparators

Zhang–Luck, fixed DNN–TCC, validation-fitted DNN–TCC, Bates-style noisy-winner TCC, and the PRISM component models all predict the identical 360-way outer-test rows (Table 7).

Table 8 reports the corrected comparison of distance plus one HOG descriptor plus stage 3, not the six-descriptor model. Its mean NLL is 4.46666 on A Walk and 4.19753 on Scene Wheels. Bounds use the crossed participant–target bootstrap (10,000 draws) jointly over eight comparators, with 97.5% simultaneous coverage per dataset and Bonferroni 95% coverage across both datasets. Only noisy-winner scores are replaced by converged 200-node quadrature scores; the entire eight-comparator family is recomputed.

Secondary metrics in Table 9 are descriptive and do not uniformly favor PRISM. Rank is the descending midrank of the observed response, with ties averaged; MRR averages its reciprocal. Top-choice expected calibration error (ECE) uses ten equal-width bins of maximum predicted probability, weighting each bin's absolute difference between mean confidence and observed top-choice

Table 7: Model parameter counts and fitting protocols.

Family	A Walk scalars	Scene scalars	Fitting before shared outer test
Zhang–Luck	342	60	Train fit, validation stopping, five starts
Fixed DNN–TCC	0	0	Fixed population constants
Fitted DNN–TCC	4	4	720-cell validation grid
Bates noisy-winner TCC	3	3	240 cells; checked by 200-node quadrature
Distance only	2	2	Estimated 10-position bin profile; scale and lapse
Distance + HOG	3	3	One nonnegative visual coefficient
Distance + final ViT	3	3	One nonnegative visual coefficient
Distance + stage 3	3	3	One nonnegative visual coefficient
HOG + stage 3	4	4	One HOG plus one stage-3 coefficient

Table 8: Held-out NLL advantage of distance plus HOG plus stage 3, with corrected one-sided simultaneous lower bounds (LB). Positive values favor this model; the noisy-winner comparator uses 200-node quadrature.

	A Walk		Scene Wheels	
Comparator	Gain	Simult. LB	Gain	Simult. LB
Zhang–Luck	.13401	.04503	.27352	.12573
Fixed DNN–TCC	1.09181	1.00284	1.25613	1.10834
Fitted DNN–TCC	.28948	.20050	.15544	.00765
Bates-style TCC	.08780	−.00118	.16364	.01586
Distance only	.08487	−.00411	.21714	.06936
Distance + HOG	.03143	−.05754	.02721	−.12058
Distance + final ViT	.01888	−.07009	.06911	−.07867
Distance + stage 3	.00164	−.08734	.01817	−.12962

accuracy by its trial share. Empty bins contribute zero. This assesses top-choice confidence, not calibration of the entire response distribution; lower ECE does not imply better NLL.

Table 9: Descriptive observed-response rank, mean reciprocal rank (MRR), and top-choice ECE on the original comparison trials.

	A Walk			Scene Wheels		
Model	Rank	MRR	ECE	Rank	MRR	ECE
Zhang–Luck	45.28	.1483	.0204	37.59	.1797	.0367
Fixed DNN–TCC	43.41	.1467	.0479	34.42	.1787	.0635
Fitted DNN–TCC	43.57	.1486	.0306	34.67	.1786	.0249
Bates-style TCC	45.72	.1538	.0123	35.93	.1781	.0316
Distance only	42.62	.1190	.0003	37.53	.1374	.0001
Distance + HOG	40.75	.1538	.0308	33.29	.1806	.0044
Distance + final ViT	40.17	.1478	.0110	34.42	.1789	.0036
Distance + stage 3	39.03	.1512	.0132	33.42	.1799	.0067
HOG + stage 3	39.10	.1530	.0264	32.66	.1802	.0071

C.6 Quadrature vs. Monte Carlo Sensitivity

Deterministic numerical quadrature leaves all eight noisy-winner parameter selections unchanged. Increasing the Gauss–Hermite quadrature resolution from 100 to 200 integration nodes changes mean held-out NLL by less than 10^{-6} nats per trial. The original Monte Carlo approximation raises its NLL by .00211 and .00234, respectively, relative to quadrature. Recomputing the complete inference family changes the noisy-winner simultaneous lower bounds from .000934 to $-.001177$ on A Walk and from .018199 to .015859 on Scene Wheels (Table 8). Thus mean NLL favors the full model on both datasets, but simultaneous superiority over this comparator is supported only on Scene Wheels. The correction does not subtract numerical error from an old confidence bound.

D PRISM-CHANNEL DETAILS AND CONTROLS

D.1 COGNITIVE BASELINE & MIXTURE MODEL

The memory baseline $q_{s,r}(c \mid t)$ models standard continuous recall through target precision $\kappa_{s,r}$ and uniform lapse $g_{s,r}$:

$$q_{s,r}(c \mid t) = (1 - g_{s,r})\,\mathrm{VM}_{360}(c - t; \kappa_{s,r}) + \frac{g_{s,r}}{360}. \tag{6}$$

Parameters are estimated via maximum likelihood on individual training splits.

D.2 VISUAL REWEIGHTING & P50 CONSTRAINT

Given the standardized mean residual similarity $z(t, c)$, the visual channel proposal is

$$v_s(c \mid t) \propto q_{s,r}(c \mid t) \exp\{T_s z(t, c)\}. \tag{7}$$

$z(t_i, c)$ is a scalar arithmetic mean over encoders, distinct from CCGD's vector $\mathbf{z}_i(c)$ of individual residual similarities (Appendix G.1). The P50 transformation operator $\mathcal{S}_{.50}(q, v)$ isolates far-tail errors: let $C_{.50}$ be the minimal target-centered candidate set containing at least 50% of baseline probability mass under q. For $c \in C_{.50}$, $\mathcal{S}_{.50}(q, v)(c) = q(c)$. For $c \notin C_{.50}$, probabilities in $v(c)$ are rescaled so that $\sum_{c \notin C_{.50}} \mathcal{S}_{.50}(q, v)(c) = \sum_{c \notin C_{.50}} q(c)$.

D.3 CROSS-FITTING PROTOCOL & PARAMETERS

Participants are calibrated on half of their trials to predict the remaining half in a 2-fold cross-validation scheme. Temperatures $T_s \in \{.25, .5, 1, 2, 4\}$ and mixing weights $a_s \in \{0, .1, .25, .5, .75, 1\}$ are selected by training-half likelihood, then fixed when scoring the opposite half. The original observer-minus-shared contrast is .00174 nats/trial (Holm $p = .529$); positive shared prediction does not require an advantage from observer-specific mixing weights. Main-paper Figure 4 compares shared-weight P50, observer-specific P50, and observer-specific unrestricted variants. Its points and numeric labels are participant means and its bars are 95% participant-bootstrap intervals; the main text separately gives the trial-weighted far-error gain. P50 preserves each central candidate probability and total far-error mass, so a far-error log-likelihood gain does not imply that far errors occur less often.

D.4 BASELINE FLEXIBILITY AND TARGET SPECIFICITY

The additional baseline checks retain the same 5,400 Scene Wheels trials, 20 observers, and opposite-half calibration as the original Channel. Each baseline uses the same 30 temperature/mixing-weight settings. Besides the original grid, continuous optimization fits a single von Mises plus uniform lapse or a two-von-Mises mixture plus lapse, with concentrations in $[.25, 256]$, lapse in $[0, .4]$, and mixture weight in $[0, 1]$. Fixed optimizer starts and all parameters use calibration responses only. The original and optimized-single parent NLLs are 4.426345 and 4.454884; the two-component parent achieves 4.317166. Thus greater flexibility does not invariably improve held-out fit, but Channel adds prediction beyond each parent (Table 10).

Table 10: Channel checks on Scene Wheels. Gains weight the 20 participants equally; intervals are pointwise participant-bootstrap 95% intervals, not simultaneous bounds. The rows use the distinct contrasts defined in the text.

Contrast	Gain (nats/trial)	95% interval
Channel − original grid parent	.037571	[.020400,.057354]
Channel − optimized single-VM parent	.041184	[.021569,.064442]
Channel − two-VM mixture parent	.043917	[.024893,.066216]
Exact-distance: Channel − grid parent	.023949	[.012752,.037952]
Original − target-perturbed Channel	.091461	[.045200,.154145]

The main text retains the original primary-analysis bootstrap interval [.02040,.05743]. The near-identical endpoint [.020400,.057354] above comes from the later common-seed, 20,000-draw bootstrap used for these additional checks; the observations and fitted models are unchanged.

Exact-distance prediction renormalizes probabilities over legal candidates at the observed absolute offset. The 371 singleton cases contribute zero conditional log score; all 5,400 trials remain in the mean. This endpoint tests identity at fixed distance, not full-distribution likelihood or P50 tail-mass conservation.

Across 20 fixed target perturbations, target rows are permuted within each wheel–radius block and candidate indices aligned to preserve signed offset; fitted weights, baseline probabilities, and P50 gates remain fixed. These are frozen-model perturbations, not refitted permutation tests. For the last table row, average the 20 perturbations within participant before resampling participants; seeds are not independent observers. Intervals use 20,000 bootstrap draws. The original single-control comparison instead reports the control's $-.00318$ gain over baseline and Channel's .04076 advantage over that control; it is distinct from the repeated-perturbation contrast. These checks support target-matched visual prediction beyond baseline spread, not a unique causal memory mechanism.

D.5 Common Task and Calibration Requirements

Each fold excludes a complete wheel and its four observers from population fitting. We evaluate identical test halves after 0, 25, 75, or 135 calibration trials per observer; nonzero budgets take the earliest 5, 15, or 27 calibration trials per radius. Both half directions are scored. Global Channel and the population baseline use no test-observer calibration. CCGD and equal-ensemble visual weights remain population fitted while their cognitive baseline adapts; observer Channel may also adapt its temperature and mixing weight.

Table 11: Common-task NLL on Scene Wheels; lower is better. Global Channel and the population baseline ignore extra calibration observations, and every column scores the same test rows.

Trials	Pop. q	Global Channel	Observer q	Observer Channel	CCGD	Equal
0	4.342244	4.312953	4.342244	4.312953	4.326520	4.326911
25	4.342244	4.312953	5.176008	5.133391	5.160280	5.160643
75	4.342244	4.312953	4.589322	4.552463	4.573225	4.573599
135	4.342244	4.312953	4.426345	4.388774	4.410331	4.410722

Shared Channel has the lowest mean NLL in this common task, supporting a population-level visual contribution without observer calibration. Individual calibration improves with sample size but does not surpass the global model. This is a calibration-burden result, not evidence that additional observations intrinsically harm prediction; hierarchical shrinkage was not tested. The held-out unit remains five wheels, not 20 independent environments.

E EEG Representational Similarity Analysis

E.1 Independent EEG Dataset & Preprocessing

The public Kiat–Luck archive provides 50-by-50 scene-similarity matrices for 29 included participants at 425 EEG time samples, along with RGB and CORnet-S controls (Kiat & Luck, 2026). The source code specifies a -500-ms origin and 4-ms sampling. We use the complete 200–996-ms window after scene onset. Because scenes were presented for 200 ms, this interval is 0–796 ms after scene offset and contains 200 time samples.

E.2 Partial RSA Pipeline

The exact behaviorally tested ViT stage-3 node, OpenCV HOG descriptor, and 32-bin-per-channel RGB histograms are applied to the 50 scenes. Pairwise model dissimilarity is one minus cosine similarity. For each participant and time sample, EEG dissimilarity is Spearman-correlated with stage-3 dissimilarity. The primary partial RSA rank-residualizes HOG and RGB from both stage-3 and EEG dissimilarities before correlation. The six-test family contains raw stage-3 RSA, partial RSA given HOG, partial RSA given RGB, partial RSA given both, stage 3 minus HOG, and stage 3 minus RGB.

Explicitly, vectorize the 1,225 upper-triangle pairs and rank each predictor and EEG vector, assigning average ranks to ties. Let X contain an intercept and the ranked HOG/RGB controls, let $M_X = I - XX^+$, and let u, v be the ranked model and EEG vectors. Partial RSA is

$$r_{\text{partial}} = \frac{(M_X u)^\top (M_X v)}{\|M_X u\|_2 \|M_X v\|_2}. \tag{8}$$

Here X^+ is the Moore–Penrose pseudoinverse and I the identity on scene pairs; the implementation returns zero if either residual norm is zero. Appendix G.5 uses the same correlation after averaging EEG similarities in time; this is not the temporal average of timewise correlations reported here.

E.3 Time Course and Participant-Level Inference

Each comparison is averaged over the fixed interval within participant and tested with 100,000 one-sided sign flips; Holm correction covers the six declared comparisons.

Table 12: Participant-level mean timewise Spearman RSA over 200–996 ms after scene onset (0–796 ms after offset). The final two rows subtract raw RSA values within participant over this same window, not partial RSA values.

Stage-3 comparison	Mean	Median	Positive	Holm p
Raw RSA > 0	.06161	.06780	27/29	6.0×10^{-5}
RSA | HOG > 0	.05777	.07124	27/29	6.0×10^{-5}
RSA | RGB > 0	.05503	.06064	27/29	6.0×10^{-5}
RSA | HOG+RGB > 0	.05394	.06351	25/29	6.0×10^{-5}
Raw stage 3 minus raw HOG > 0	.02638	.03379	20/29	.032540
Raw stage 3 minus raw RGB > 0	.01589	.01997	18/29	.034480

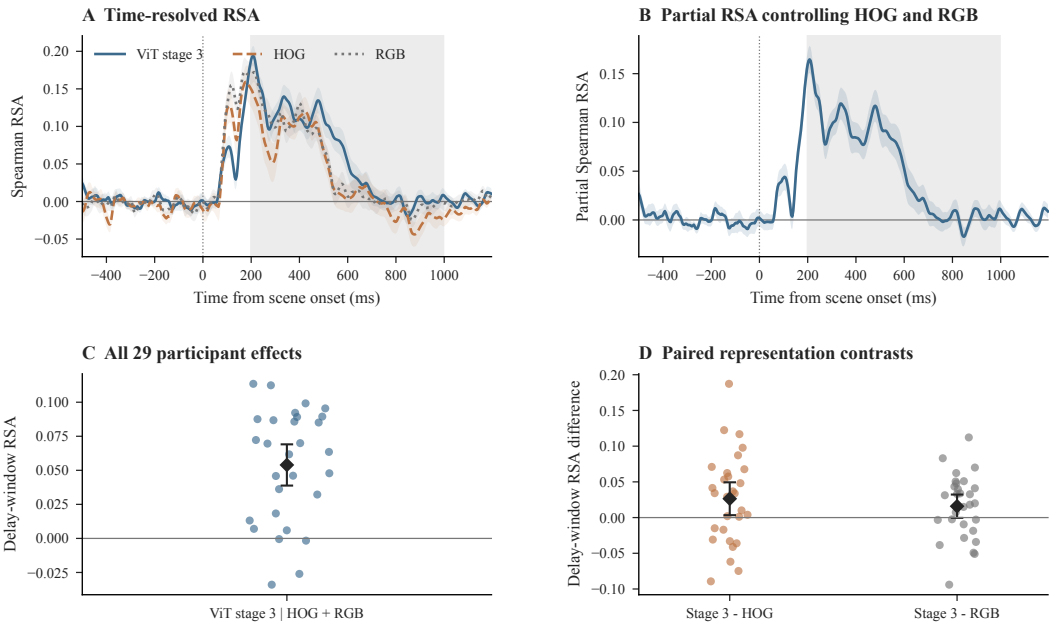

Figure 1: **Sensor-space EEG correspondence. A:** Time-resolved RSA for ViT stage 3, HOG, and RGB; shading is participant SEM. **B:** Stage-3 partial RSA after HOG+RGB removal; gray marks the 200–996-ms analysis window. **C:** All 29 participant effects. **D:** Paired stage-3-minus-control effects. Diamonds in C–D show means with descriptive normal-approximation intervals (± 1.96 SEM); inference uses the sign-flip tests in Table 12. The analysis tests shared representational geometry, not anatomical localization or causality.

F Nested Cross-Validation & Model Selection

For each held-out route or wheel, PRISM-Eval selects representations strictly within the remaining training environments; no held-out response enters model selection. The complete target-disjoint procedure is:

1. Training environment trials are split into 5 non-overlapping target folds.
2. The regularization grid $\xi \in \{0.001, 0.01, 0.1\}$ and encoder/layer candidates are evaluated via 5-fold cross-validation.
3. Optimal parameters are refit on the full training set before scoring the held-out test environment once.

CCGD's primary inference uses the nested ridge comparison in Appendices G.2 and G.6–G.7, which holds out development environments, not five target folds. The original fixed-penalty analysis is retained as a historical exploratory benchmark.

G Cognitive-Channel Geometry Distillation (CCGD)

G.1 Mathematical Formulation & Convexity

CCGD combines $K = 3$ frozen visual similarity fields $\mathbf{z}_i(c) = [z_{i,1}(c), z_{i,2}(c), z_{i,3}(c)]^\top$ (representing residual similarities from ViT, ResNet, and ConvNeXt stage 3) with an environment-excluded cognitive memory baseline $q_i(c) > 0$. The predicted probability of choosing candidate $c \in \{0, \ldots, 359\}$ on trial i is

$$p_{\mathbf{w}}(c \mid i) = \frac{q_i(c) \exp\{\mathbf{w}^\top \mathbf{z}_i(c)\}}{\sum_{d=0}^{359} q_i(d) \exp\{\mathbf{w}^\top \mathbf{z}_i(d)\}}. \tag{9}$$

The objective function over training set $\mathcal{T}$ is the regularized negative log-likelihood:

$$\mathcal{L}(\mathbf{w}) = -\frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \log p_{\mathbf{w}}(y_i \mid i) + \frac{\rho}{2} \|\mathbf{w}\|_2^2. \tag{10}$$

The gradient with respect to $\mathbf{w}$ is

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \big(\mathbb{E}_{p_{\mathbf{w}}(\cdot \mid i)}[\mathbf{z}_i] - \mathbf{z}_i(y_i) \big) + \rho \mathbf{w}, \tag{11}$$

and the Hessian matrix is

$$\nabla^2_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \mathrm{Cov}_{p_{\mathbf{w}}(\cdot \mid i)} [\mathbf{z}_i] + \rho I_3. \tag{12}$$

Because the covariance matrix of any random vector is positive semi-definite and $\rho I_3 \succ 0$ for every tested $\rho > 0$, the Hessian is strictly positive definite ($\nabla^2_{\mathbf{w}} \mathcal{L}(\mathbf{w}) \succ 0$) everywhere. Thus, $\mathcal{L}(\mathbf{w})$ is strictly convex and possesses a unique global minimum; deterministic L-BFGS-B is used with the tolerances below.

G.2 Optimization & Regularization (ρ)

Optimization uses deterministic L-BFGS-B from zero weights, with objective tolerance 10^{-12}, gradient tolerance 10^{-8}, and at most 200 iterations. The original analysis fixed $\rho = 1$ during development using both behavioral archives, before the later inner-environment grid searches. We retain it as a historical exploratory benchmark. Primary inference now uses nested selection of $\rho \in \{.01, .1, 1\}$ separately within each outer training set (Appendices G.6–G.7). This controls fold-level penalty selection but does not undo earlier specification development on these archives. Standardizing residual similarities $\mathbf{z}_i(c)$ to zero mean and unit variance places the predictors on a common scale for the isotropic penalty.

G.3 Baseline & Out-of-Fold Prediction

For A Walk in the Park, leave-one-route-out cross-validation fits CCGD on two routes and evaluates on the third withheld route. For Scene Wheels, leave-one-wheel-out fits CCGD on four wheels and evaluates on the fifth held-out wheel alongside four unseen participants.

G.4 Participant-Level & Environment Generalization

Table 13 reports environment-level baseline gains under the fixed $\rho = 1$ specification across outer test folds; these environment breakdowns serve as descriptive evaluations of the fixed pipeline prior to nested selection.

Table 13: Original fixed-$\rho = 1$ CCGD: trial-weighted held-out log-score gains (nats/trial) across environments.

Dataset	Environment	CCGD − baseline	CCGD − equal
A Walk	JMP3	.016903	.000320
	OLM7	.012513	.000603
	STB9	.012726	.000891
Scene Wheels	1	.013228	−.000396
	2	.001753	.000172
	3	.013926	.000764
	4	.019130	.000785
	5	.030582	.000629

At the individual participant level, CCGD yields mean participant log-likelihood gains of 0.01301 nats/trial (95% CI $[0.01056, 0.01526]$, $p \leq 10^{-6}$, 110/114 positive) on A Walk in the Park and 0.01572 nats/trial (95% CI $[0.01075, 0.02115]$, $p = 7.63 \times 10^{-6}$, 18/20 positive) on Scene Wheels. A Walk's bound is the resolution-limited Monte Carlo estimate $1/(1{,}000{,}000 + 1)$, not an exact tail probability; the Scene Wheels test enumerates participant sign patterns.

Table 14 gives all-behavior fits used for EEG transfer; behavioral test scores instead use their outer-fold fitted weights. These coefficients multiply standardized visual predictors in conditional log odds. They are not mixture proportions and need not sum to one. Because the predictors are correlated and the readout is regularized around a cognitive baseline, individual coefficients depend on the other predictors, the penalty, and that baseline; they are not measurements of brain-region activity.

Table 14: All-behavior CCGD weights and their equal-dataset average for EEG transfer.

Fit	ViT stage 3	ResNet50 stage 3	ConvNeXt-B stage 3
A Walk	.080906	.043003	.048505
Scene Wheels	.081982	.070472	.048878
Equal-dataset average	.081444	.056738	.048691

G.5 EEG Cross-Modal RSA Alignment

The equal-dataset average weights in Table 14 combine the three raw cosine-similarity matrices of the 50 EEG scenes, without fitting to EEG or applying route/wheel offset residualization. The equal-weight ensemble averages the same matrices. For each of 29 participants, partial Spearman RSA uses all 1,225 unordered scene pairs, controlling HOG and RGB.

Here EEG matrices are averaged over 200–996 ms before partial RSA is computed, whereas Appendix E computes RSA at each time point and then averages the correlations. Because correlation and temporal averaging do not commute, the stage-3 references (.05511 here and .05394 in the timewise analysis) estimate different quantities.

CCGD exceeds stage 3 by .00516 (participant-bootstrap 95% interval [.00027,.00993], nominal one-sided $p = .02489$; 20/29 participants positive). The equal-weight ensemble has slightly higher mean

Table 15: Partial RSA of delay-averaged EEG matrices, controlling HOG and RGB.

Visual geometry	Mean partial RSA
ViT-B/16 stage 3	.055111
CCGD, behavior-trained	.060268
Equal-weight three-backbone ensemble	.061627

RSA than CCGD; these results support cross-modal visual structure, not superior neural alignment of learned unequal weights.

G.6 Equal-Weight Control Comparison

To separate relative-weight learning from ensembling, CCGD is compared against a constrained equal-weight control $z_{\text{eq}} = (z_1 + z_2 + z_3)/\sqrt{3}$. The primary inference uses the separately nested comparison in Appendix G.7. CCGD minus the fitted equal-weight control is .000516 nats/trial on A Walk (participant-bootstrap 95% interval [−.001607,.003016]) and −.002267 on Scene Wheels ([−.004991,.000093]). Both intervals include zero: an unequal-weight advantage is not established under nested selection. This does not establish equivalence or equal-ensemble superiority.

For the historical fixed-$\rho = 1$ benchmark, the participant-mean increments are 0.000785 nats/trial on A Walk (original one-sided $p \leq 10^{-6}$, simulation-limited; 90/114 positive) and 0.000391 on Scene Wheels (original one-sided $p = .00771$, 16/20 positive). Transforming a mean log-score gain Δ as $100(\exp(\Delta) - 1)$ gives .079% and .039% increases in geometric-mean observed-response probability, not accuracy percentage points. These exploratory results and their original tests are distinct from the expanded diagnostic inference below.

The matched A Walk fixed-penalty readout uses training-fitted rather than cross-fitted training q and gives .000793 [.000535,.001111]; its difference from the original increment is only .00000814. Thus the nested comparison is not explained solely by parent construction. In the eight-contrast diagnostic family, two-sided participant sign flips with Holm correction give $p = 1.000$ and .3295 for the nested increments, .0788 for fixed Scene Wheels, and 4.0×10^{-5} for both original and matched fixed A Walk. The Scene Wheels nested-minus-fixed change is $-.002658$ [−.005256,−.000429], but its adjusted $p = .1858$. Intervals are pointwise 95% participant bootstraps (100,000 draws); tests use 200,000 sign flips. These are new inference summaries of retained scores, not refitted models, and condition on the observed environments.

G.7 Alternative Cognitive Baselines

The primary nested matched comparison uses identical outer-test rows, candidate sets, gallery preprocessing, and log scores. Within each outer training set, leave one development environment out at a time (two inner routes or four inner wheels). For each readout family, mean inner-environment NLL selects $\rho \in \{.01, .1, 1\}$, with ties favoring the larger penalty, before refitting on all outer-training trials. CCGD fits three coefficients; equal weighting fits one common strength with matching penalty on the equal-coefficient subspace. Single-backbone selection considers nine backbone/ridge pairs; HOG is refitted rather than borrowed from PRISM-Eval.

A Walk's flexible baseline adds four signed-distance polynomial coefficients with fixed ridge .1; Scene Wheels uses two von Mises concentrations, a mixture weight, and a lapse per radius. All fits exclude test responses. The new A Walk readout uses training-fitted q, whereas the original fixed-penalty readout used cross-fitted training q; their difference is not a penalty-only effect. Adding CCGD to the flexible baseline gives participant-mean gains of .018075 [.013556,.022442] on A Walk and .028015 [.010582,.046919] on Scene Wheels (pointwise 95% participant-bootstrap intervals). Uniform $q = 1/360$ yields poor visual-only prediction. These results support cognitive/topological conditioning and an additional visual contribution, not the unique validity of one baseline family or a general advantage for unequal weighting.

Table 16: Separately nested matched readouts: trial-weighted held-out NLL, lower is better. These are distinct from the original fixed-penalty results.

Model	A Walk	Scene Wheels
Cognitive q	4.597960	4.342244
Selected single backbone + q	4.573649	4.324462
Fitted equal ensemble + q	4.577293	4.315819
CCGD + q	4.578402	4.318086
HOG + q	4.592051	4.317143
Flexible q	4.576212	4.262854
CCGD + flexible q	4.556889	4.234839
Uniform q	5.886104	5.886104
CCGD + uniform q	5.877068	5.878483

G.8 Exploratory Fixed-Penalty Checks: Detrending & Spatial Spacing Sensitivity

With ridge fixed at 1, we compare signed-offset centering with degree-2 and degree-4 polynomial removal from the response-free gallery similarity matrices. For A Walk, we also add neighboring-image similarity at lags 1, 2, and 4, alone or with the original three visual covariates. These are visual spatial controls, not nonvisual baselines or proof of independent adjacent frames.

Table 17: Participant-mean visual gains over the cognitive baseline under detrending and spatial controls, with pointwise 95% bootstrap intervals.

Variant	A Walk	Scene Wheels
Signed-offset centering	.013015 [.010538,.015258]	.015724 [.010784,.021156]
Degree-2 detrending	.015653 [.003507,.026411]	.016255 [−.001566,.031404]
Degree-4 detrending	.003322 [−.004299,.010058]	.019627 [−.000380,.036569]
Spatial-lag control	.011574 [.009263,.013644]	Not evaluated
CCGD + spatial control	.015796 [.012454,.018833]	Not evaluated

Conditioning on the observed absolute offset and renormalizing over legal equal-distance candidates gives CCGD-minus-baseline gains of .002061 [.000131,.003625] on A Walk and .009142 [.006428,.012037] on Scene Wheels. The corresponding gains over the equal-weight model are .000222 [.000083,.000393] and .000162 [.000011,.000305]. Intervals are pointwise participant-bootstrap 95% intervals; singleton offsets contribute zero.

Wider candidate spacing provides a complementary sensitivity test. Strides 2 and 4 retain trials only when both target and observed response lie on the coarsened grid, and every comparator uses the same reduced choices. The response-based eligibility changes the task; raw NLL must not be compared across grids or treated as an unbiased replacement for the 360-way result.

Under the exploratory fixed-$\rho = 1$ setting, the visual increment and a small unequal-weight increment survive both candidate-spacing settings; however, as shown in Appendix G.6, the unequal-weight advantage does not generalize under strict nested selection. Joint visual and spatial controls also retain positive A Walk prediction. Stronger polynomial detrending is less conclusive: the quartic A Walk interval and both Scene Wheels polynomial intervals include zero. These checks support specific forms of spatial robustness, not complete removal of topology dependence.

Table 18: A Walk coarsened-task sensitivity. Gains and intervals weight participants equally; retained/excluded counts refer to the original 15,035 trials.

Stride	Candidates	Retained/excluded	CCGD − q	CCGD − equal
2	180	3790/11245	.017000 [.012637,.022347]	.000302 [.000045,.000571]
4	90	967/14068	.022690 [.015769,.029625]	.001006 [.000440,.001592]

H Data Access, Preprocessing, and Licensing

All analyses rely on public open-access archives:

- **A Walk in the Park:** Data and materials from Westebbe et al. (2024) are available at `https://osf.io/a7mkt/`. The OSF project metadata does not specify a dataset license; raw data and row-level caches are not redistributed here.
- **Scene Wheels:** Data and stimulus continua were released by Son et al. (2022) at `https://osf.io/h5wpk/` under CC BY 4.0. Bates et al. (2024) provides the related computational modeling study.
- **EEG Archive:** Data provided by Kiat & Luck (2026) on OSF.

I Additional Architecture Experiments

EfficientNet-B0 was absent from pipeline development and evaluated only after the public encoder interface and analysis were fixed. This is a post-pipeline portability demonstration, not independent scientific confirmation. Adding its frozen representation beyond response distance and HOG improves held-out NLL in both datasets (Table 19).

Table 19: EfficientNet-B0 gain beyond response distance and HOG. Intervals use the crossed participant–target bootstrap with 10,000 draws.

Dataset	Trials	Participants	Mean gain	95% CI
A Walk	15,035	114	.03709	[.02932,.04476]
Scene Wheels	5,400	20	.01756	[.00934,.02563]

Every route and wheel is positive: the A Walk gains are .04110, .03009, and .04009 for JMP3, OLM7, and STB9; the Scene Wheels gains are .01645, .00917, .02038, .01698, and .02483 for wheels 1–5. The unchanged contract validator also rejects checksum mutation, candidate-row reordering, a response-bearing array, and a zero-norm embedding.

J Implementation and Data Flow

J.1 Cognitive Baselines and Calibration

The A Walk CCGD baseline is not PRISM-Eval's empirical distance profile. For $d_i(c) = c - t_i$, its initial mixture is

$$q_i^0(c) = A\frac{\exp[-(d_i(c) - B)^2/(2S^2)]}{\sum_{j=0}^{359}\exp[-(d_i(j) - B)^2/(2S^2)]} + \frac{1 - A}{360}. \tag{13}$$

The three parameters are shared across training observers and routes: $A = .05 + .90\,\mathrm{sigmoid}(a)$, $B = 20\tanh(b)$, and $S = 1 + 79\,\mathrm{sigmoid}(s)$. Normalization uses the legal finite candidate set for each target, without wrapping or treating nonexistent route positions as possible reports. A three-coefficient conditional extension gives $q_i(c) \propto q_i^0(c)\exp\{\boldsymbol{\gamma}^\top \boldsymbol{\psi}_i(c)\}$; its fit minimizes mean training NLL plus $\|\boldsymbol{\gamma}\|^2/2$. The resulting q is held fixed when fitting the visual readout.

The tangent features $\boldsymbol{\psi}$ are constructed as follows. Compute the three candidate-wise score derivatives of $\log q^0$ with respect to availability, bias, and log scale, and center them under q^0. Multiply these scores by a target-repetition conditioner: $\log(1 + n_{t_i})$, where n_t counts occurrences of target t in the fixed structural trial table. These counts use target identifiers, not reports or response accuracy. The retained conditioner is globally standardized, then re-centered and scaled using only each fit's recorded training trials before multiplication; the latter standardization removes the initial affine scaling. This is the actual conditioner in the retained parent, not an image-density covariate. Under the training inner product $\langle u, v\rangle = |\mathcal{T}|^{-1}\sum_{i\in\mathcal{T},c} q_i^0(c)u_i(c)v_i(c)$, project the conditioned scores off the unconditioned score span using the Gram-matrix pseudoinverse. Center the residuals under q^0 and Fisher-whiten their three coordinates, rejecting eigenvalues $\le 10^{-10}$. Freeze this transform for validation and test trials. This defines a conditional baseline, not an additional visual-encoder fit.

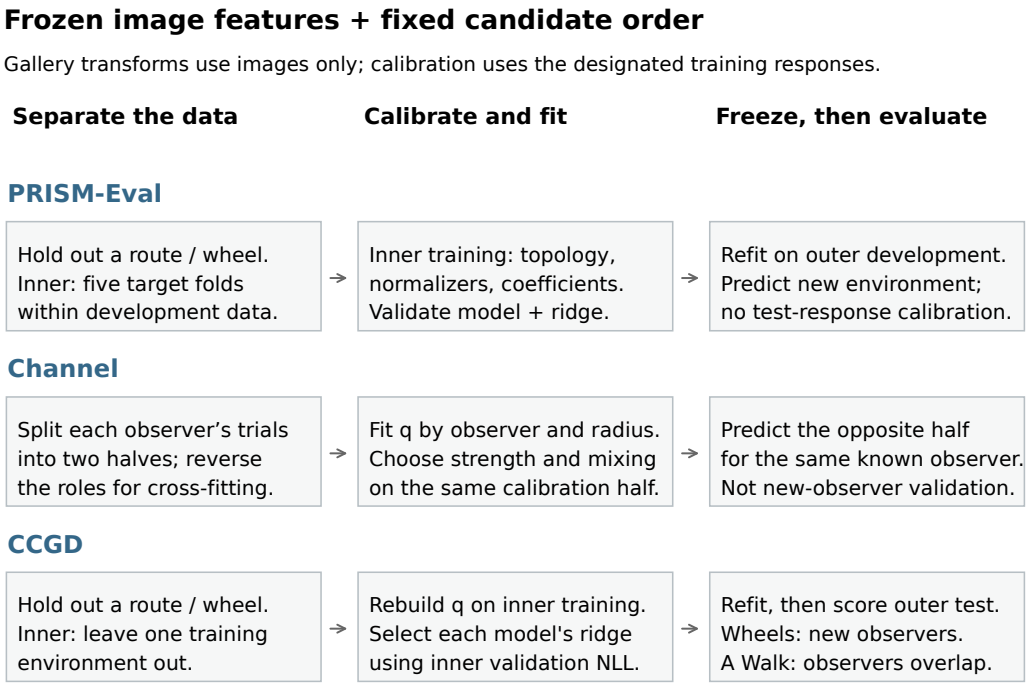

Figure 2: Distinct fitting and evaluation paths. The top row expands main Figure 3A; that target-fold scheme does not describe Channel or CCGD. Response-derived quantities are reconstructed on the indicated training responses. Channel and CCGD offset-centering and residual scaling use all image pairs in a candidate gallery, including a test gallery, but no memory reports. The CCGD row shows primary nested selection; its historical fixed-penalty construction is described below.

Original fixed-penalty A Walk CCGD uses the retained cross-fitted q values for training responses and a parent fitted without the test route for test responses. The matched fixed and nested comparisons instead fit the parent on each relevant training set and use its training-fitted probabilities for the readout. The matched-fixed comparison isolates this change from ridge selection (Appendix G.6). In nested fits, the mixture, tangent transform, and conditional coefficients are rebuilt inside each inner training set; no inner-validation response fits its own parent. Outer refitting repeats this construction on all outer-training environments. A Walk observers can occur in both training and held-out routes; these tests do not withhold observers.

For Scene Wheels, CCGD fits the discrete von Mises–uniform baseline in Appendix D.1 separately by radius, pooling training observers. The concentration grid is $\{.25, .5, 1, 2, 4, 8, 16, 32, 64, 128\}$ and the guessing grid is $\{0, .01, .025, .05, .1, .2, .4\}$ (70 likelihood-evaluated pairs). Original Channel uses the same grid for each observer and radius on the opposite trial half, followed by 30 strength/mixing pairs from Appendix D.3. The common-task comparison instead separates population fitting from 0/25/75/135 test-observer calibration trials (Appendix D.5). Fixed CCGD estimates three weights at $\rho = 1$; nested CCGD and the one-strength equal control each search three penalties using the same inner-environment tests. Thus these protocols have different calibration budgets, not interchangeable definitions of a held-out observer.

J.2 Frozen Stage Features

The four sampled ResNet50 stages are the final residual outputs of layers 1–4 (torchvision nodes `layer1.2.relu_2`, `layer2.3.relu_2`, `layer3.5.relu_2`, `layer4.2.relu_2`); each is spatially averaged. ConvNeXt-B uses the final residual additions of its four stages (feature nodes `1.2.add`, `3.2.add`, `5.26.add`, `7.2.add`), also spatially averaged. ViT-B/16 uses

the CLS token after blocks 3, 6, 9, and 12 (zero-based encoder-layer indices 2, 5, 8, 11, `add_1`). Stage 3 therefore has a fixed architectural meaning for each network, not a layer chosen separately for individual test responses.

The torchvision checkpoint enums are `ResNet50_Weights` with `IMAGENET1K_V2`, `ConvNeXt_Base_Weights` with `IMAGENET1K_V1`, and `ViT_B_16_Weights` with `IMAGENET1K_V1`. RGB images are resized by their shorter side to 232, 232, and 256 pixels, respectively, with bilinear antialiasing, then center-cropped to 224^2. Channel means $(.485, .456, .406)$ and SDs $(.229, .224, .225)$ standardize inputs. The retained extraction environment uses PyTorch 2.9 and torchvision 0.24. Spatial averages and CLS vectors are taken before the classification head; cosine similarity normalizes the extracted vector, not the image.

J.3 Matched Cognitive Comparator Specification

Let $S_i(c)$ be the fixed target–candidate cosine similarity and let $D_i(c) = |c - t_i|/359$ for A Walk or circular shortest distance divided by 180 for Scene Wheels. The DNN–TCC comparator uses $p_i = (1 - g)\,\mathrm{softmax}((\alpha S_i - \beta D_i)/T) + g/360$. The fixed version has $(\alpha, \beta, T, g) = (1, .1, 1, 0)$. The fitted version searches $\alpha \in \{.25, .5, 1, 2, 4, 8\}$, $\beta \in \{0, .01, .03, .1, .3, 1\}$, $T \in \{.5, 1, 2, 4\}$, and $g \in \{0, .02, .05, .1, .2\}$ (720 combinations). These are comparator implementations, not the claim that softmax is the same probability construction as the original TCC model.

The noisy-winner comparator selects $\arg\max_c\{d' S_i(c) - bD_i(c) + \epsilon_c\}$ with independent $\epsilon_c \sim \mathcal{N}(0, 1)$, followed by uniform mixing with probability g. Its grid is $d' \in \{.25, .5, 1, 1.5, 2, 3, 4, 6\}$, $b \in \{0, 2, 4, 8, 16, 32\}$ and the same five g values (240 combinations). Both grids minimize validation NLL, with the recorded candidate order breaking ties, and share one parameter set across a fold rather than fitting each test observer. Winner probabilities are evaluated by one-dimensional Gaussian quadrature: originally 25 nodes for validation and averages of five 4,096-draw Monte Carlo estimates for test probabilities, now checked with 200-node quadrature throughout. The corrected score-family inference is in Appendix C.5; all grid selections remain unchanged. The separate participant-parameter Zhang–Luck comparator is not the shared A Walk conditional baseline defined above.

Compact fitting algorithm. (1) Fix images, candidate order, features, and outer test identifiers. (2) Within each declared training split, construct response-derived topology or cognitive parameters and their training-only transforms; compute response-free gallery similarities separately. (3) Fit coefficients and, where specified, select hyperparameters only on the corresponding inner validation data. (4) Reconstruct the selected model on the full outer training set, freeze it, and score each outer test response once. (5) Retain paired observed-response log scores with participant, target, and environment identifiers for the declared resampling analysis. An automated dependency audit replayed all 80 retained model selections and six A Walk parent fits. Recorded inner and outer environment splits were disjoint, selections reproduced from inner losses, and no test-response dependency was detected in the checked fitting operations. These checks validate the replayed records, not unexamined historical analyses.

K Sensitivity, Broader Controls, and Transfer

K.1 Parameter Sensitivity

Seven one-factor-at-a-time PRISM-Eval settings retain the nominated ViT stage-3 representation: lapse caps $.1, .3, .5$; distance-bin widths 5, 10, 20; and pseudocounts $.1, 1, 10$. The shared reference setting is $(.3, 10, 1)$. Within each original outer fold, the six-descriptor baseline selects its ridge on the training/validation partition; both models are then refitted on training plus validation and scored on the same outer test responses. This is a fixed-representation sensitivity analysis, not another search over encoders or a choice of the best sensitivity setting.

All seven gains beyond the joint six descriptors are positive in both datasets (Figure 3A–C): .02675–.02941 nats/trial on A Walk and .01780–.01992 on Scene Wheels. Within each dataset, max-deviation lower bounds use 10,000 crossed participant–target bootstrap draws and a .025 family error allocation for the seven settings; the two datasets together have a Bonferroni .05 allocation.

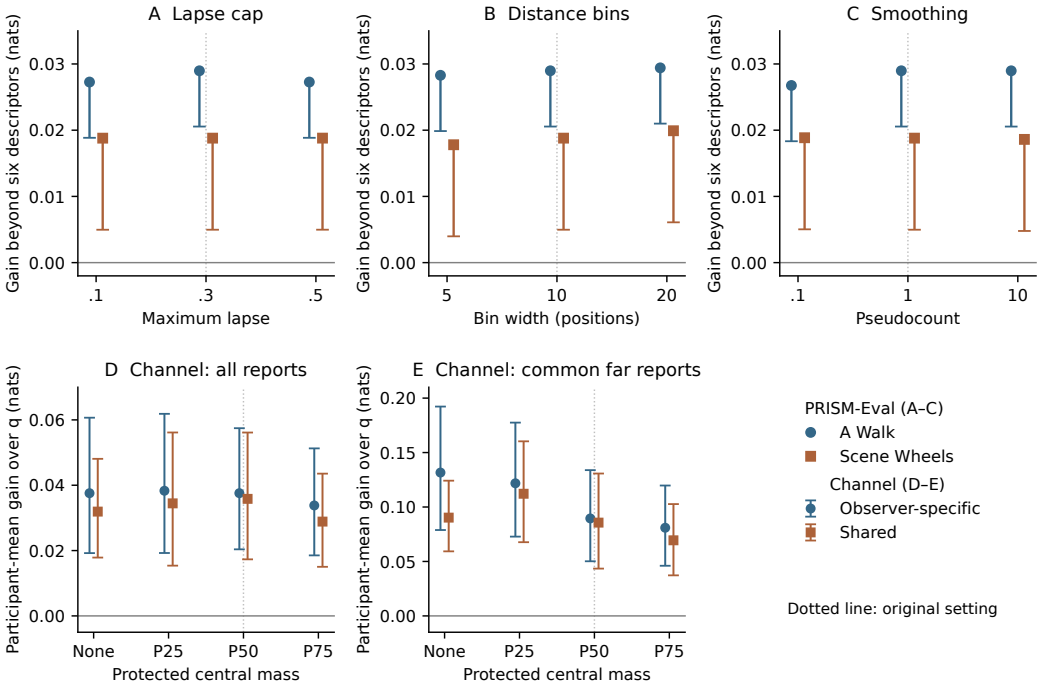

Figure 3: Parameter sensitivity with the original setting marked by a dotted line. **A–C:** Trial-weighted held-out stage-3 gain beyond six jointly fitted descriptors; each segment extends from its simultaneous one-sided lower bound to the estimate. **D–E:** Participant-mean Channel gain over the cognitive baseline, with pointwise two-sided 95% intervals. Panel E scores a common baseline-defined P50 far-error set for all settings. No points are interpolated, and no setting is selected from these results.

The smallest lower bounds are .01833 and .00397. All 280 inner/final fits converge, and none of the 112 final fits lies within .001 of its lapse cap.

Channel evaluates no protection, P25, P50, and P75 with both shared and observer-specific visual parameters. Every setting retains the original opposite-half calibration. All-report gains range from .02887 to .03831 nats/trial. For comparability, every far-error test uses the same original baseline-defined P50 set, not a different tail for each protection level. All eight all-report and eight far-error contrasts against q are positive after Holm correction across 16 two-sided participant sign-flip tests (200,000 draws; largest adjusted $p = .00288$). The plotted intervals are pointwise 95% participant-bootstrap intervals (100,000 draws), not simultaneous intervals. At P50, observer-specific and shared far-error participant means are .08952 [.05017,.13387] and .08567 [.04344,.13075], respectively. Across protection settings, observer-specific mixing reaches a grid endpoint in 22–23 of 40 fits and temperature reaches an endpoint in 4–8; these are sensitivity results for the tested grid, not evidence that an unrestricted optimum has been found.

K.2 Six-Descriptor Behavioral and Neural Controls

Behavioral readouts share the same cognitive q, outer rows, and inner environment folds. Each searches $\rho \in \{.01, .1, 1\}$ using mean inner-environment NLL. The six-descriptor model learns six signed conditional-logit coefficients; the joint model adds the three neural similarities, learning nine. Table 20 reports all six primary contrasts. The joint nine-versus-six gain supports a neural contribution beyond these descriptors in both datasets; it does not imply that the three-coefficient CCGD alone outperforms the six-descriptor readout on both datasets. Intervals condition on the retained environments.

The EEG extension uses the same 29 observers, 50 scenes, and 1,225 unique scene pairs as Appendix G.5. It residualizes rank-transformed model and delay-averaged neural dissimilarities against

Table 20: Broader behavioral controls: participant-mean log-score gains, pointwise 95% bootstrap intervals (100,000 draws), and two-sided sign-flip tests (200,000 draws), Holm-adjusted across the six rows.

Dataset	Contrast	Gain	95% CI	Holm p
A Walk	Six descriptors $-q$	.00182	[−.00228,.00528]	.3607
	CCGD − six	.01644	[.01288,.02011]	3.0×10^{-5}
	Nine − six	.01189	[.00943,.01417]	3.0×10^{-5}
Scene Wheels	Six descriptors $-q$	.02894	[.01298,.04606]	.00914
	CCGD − six	−.00479	[−.00947,.00013]	.1432
	Nine − six	.00474	[.00141,.00839]	.04461

all six fixed descriptors and an intercept. This delay-averaged analysis is distinct from the timewise RSA contrasts in Table 12. The six-control design is full rank (7 columns; condition number 85.9); adding RGB gives 8 full-rank columns and condition number 87.4. For raw cosine geometries, mean partial RSA is .05374 for stage 3, .05877 for CCGD, and .06010 for equal weighting. Each is positive in 10,000 scene-label permutations (Holm $p = 3.0 \times 10^{-4}$ across the three geometries, at the simulation resolution). Thus the six descriptors do not exhaust neural alignment. The CCGD-minus-equal contrast is $-.00132$ [$-.00268$,.00008]; across raw cosine, gallery-pair standardization, and double-centering/standardization, paired intervals include zero, both with six controls and with RGB added. No scaling rule is chosen using EEG outcomes, and these tests do not establish a neural advantage for learned relative weights.

Population-vector comparison. Using the published CORnet-S V1 and IT geometries with matched scenes and six-control partial RSA gives means .00004 and .04447. CCGD minus IT is .01431 [$-.00498$,.03344] and equal weighting minus IT is .01563 [$-.00418$,.03522]; neither establishes an advantage (both Holm $p = .551$ in the eight-contrast family spanning two control sets). The positive IT mean is a descriptive convergence across visual geometries, not a significant ranking of PRISM against a population-vector theory. A noisy-vector model perturbs a neural representation and decodes it; PRISM evaluates conditional probabilities over reports. The source account does not uniquely specify a likelihood over these 360 candidates. An isotropic vector-noise/nearest-candidate adapter passed normalization, zero-noise-winner, and translation-invariance checks, but adds assumptions and was not used as a canonical behavioral benchmark.

K.3 Linear Channel and Perceptual Transfer

Channel also applies to A Walk's linear candidate set, using its normalized conditional baseline (Appendix J.1), absolute linear distance for protected regions, and no circular wrap. Training routes choose strength and mixing; the held-out route supplies no calibration responses. There are 114 participants, 15,035 reports, and 6,689 reports in the common P50 far set. Unprotected Channel gains .01680 [.01048,.02252] nats/trial over q; P50 gains .01089 [.00281,.01763]. The corresponding far-error gains are .05060 [.04176,.05881] and .03007 [.01967,.03967]. All four are positive under the eight-contrast Holm family (all-report P50 $p = .00910$; the other three $p = 4.0 \times 10^{-5}$, simulation-limited). Against the fixed equal ensemble, unprotected Channel gains .00457 [.00083,.00794] overall and .02565 [.02042,.03051] on far reports; P50's respective gains are $-.00134$ [$-.00691$,.00323] and .00512 [$-.00174$,.01141]. Thus linear transfer supports visual reweighting without requiring P50. All estimates here weight participants equally; intervals use 100,000 participant bootstraps and tests 200,000 two-sided sign flips.

Perceptual transfer is a separate task, not a memory-generalization test. Reusing the behavior-only stage-3 coefficients with raw cosine similarity gives 68.208% human-choice agreement on a fixed 1,200-triplet BAPPS validation subset (200 per distortion group), versus 68.158% for equal weighting, and 91.5% for both on the first 200 NIGHTS test reference IDs. The learned-minus-equal differences are .050 percentage points [$-.450$,.533] and .000 [-1.500,1.500], respectively. These bounded samples do not establish a relative-weight advantage in perceptual choice; they are neither full leaderboards nor the original topology-residual memory predictor. Bootstrap intervals treat selected triplets as independent (stratified by distortion for BAPPS); dependencies between source images may reduce effective sample size. No perceptual labels fit the coefficients.

K.4 Adapted Transfer to an Independent Memory Task

The published object–scene memory task of Pettini et al. (2026) provides a distinct stimulus set and author-filtered participant sample. Participants see a target for 2 s and, after an 8 s delay, choose between target and foil scenes. The archive contains 240 participants, 108 object–scene families across six categories, and 432 task JPEGs. Of 25,920 response rows, 387 lack a choice; the primary nonrepeated condition contains 12,765 valid reports. This is transfer to an existing independent study, not newly recruited participants; cross-study identity overlap cannot be established from anonymous identifiers.

The fixed source coefficients for ViT, ResNet, and ConvNeXt are $(.081444, .056738, .048691)$; normalizing their sum gives $(.435825, .303615, .260559)$. Their relative values are not relearned on the new task. Each image pair supplies the three cosine distances $1 - \cos(\phi_m(\text{target}), \phi_m(\text{foil}))$. An adapted logistic cognitive baseline includes the full interaction of intercept/centered block, three difficulty levels, and repetition status, plus target-first and correct-key-left indicators (14 columns). Block is centered/scaled as $(b - 2.5)/2.5$. The baseline uses ridge 10^{-4} on non-intercept coefficients. Within each training set, project each visual distance off this baseline design and divide its residual by its training SD (floor 10^{-8}). The transferred weighted score and equal-weight score each fit one nonnegative scalar strength under the same logistic offset and $10^{-4}\beta^2/2$ penalty. Thus the test transfers relative visual weights, with matched task-specific calibration; it is not zero-shot use of the original 360-way model or P50.

Five participant folds crossed with five object-family folds give 25 tests. Participants are assigned round-robin by sorted anonymous identifier; object families are sorted and assigned within category. For each test, training excludes both its participant fold and its object-family fold; their intersection is the test set. Every valid report is scored once. The primary estimand is the equally weighted participant-mean gain on nonrepeated trials. Table 21 gives all three primary contrasts. Crossed participant–object bootstrap intervals use 20,000 draws of retained predictions, not refitted models; two-sided participant sign flips use 200,000 draws and Holm correction across three contrasts. The positive learned-minus-equal result therefore generalizes the transferred relative weights to this adapted memory-prediction task.

Table 21: Adapted independent-memory transfer: primary participant-mean log-score gain, crossed 95% interval, and Holm-adjusted sign-flip result. All three tests reach the 1.5×10^{-5} adjusted simulation resolution.

Contrast	Gain	Crossed 95% CI	Holm p
Transferred CCGD $-q$	.008411	[.003664,.013497]	1.5×10^{-5}
Equal weights $-q$	.007811	[.003193,.012716]	1.5×10^{-5}
Transferred CCGD $-$ equal	.000600	[.000140,.001083]	1.5×10^{-5}

This contrast does not establish superiority to every individual encoder: descriptive NLL is .480844 for transferred CCGD, .481444 for equal weighting, and .480623 for ViT alone. The analysis uses the released author-filtered responses and the actual task JPEGs. Public data do not reconstruct all original exclusion decisions; historical response metadata and current task schedules differ, so original responses were retained rather than silently remapped. An outcome-independent JPEG/PNG content check supports close image correspondence, but cannot prove historical pixel identity. These provenance limits qualify transfer beyond the observed archive.

K.5 Synthetic Mechanism Recovery

The simulation tests whether crossed manipulations could discriminate two specified mechanisms; it does not identify a mechanism in the human data. Each synthetic gallery has 360 random unit vectors in eight dimensions and 12 targets spaced from positions 20 to 340. A two-by-two design crosses delay $d \in \{0, 1\}$ and response context $k \in \{0, 1\}$, with 64 reports per target and condition: 3,072 training and 3,072 independent test reports per replication. The baseline is normalized over legal candidates with log score $-\frac{1}{2}[(c - t)/(20 + 10d)]^2 - .1k \max(c - t, 0)/359$. Standardized visual similarity $z(t, c)$ enters as $(\beta_0 + \beta_D d + \beta_C k)z(t, c)$ above this offset. The representation account sets $\beta_C = 0$; the readout account sets $\beta_D = 0$. A joint model estimates all three. Fits

minimize mean NLL with ridge 10^{-6}; L-BFGS-B uses at most 200 iterations, objective tolerance 10^{-12}, and gradient tolerance 10^{-8}.

Table 22: Synthetic recovery in 100 replications per case. Selection uses independent test NLL. Under the no-modulation case neither mechanism is true.

Generating account	$(\beta_0, \beta_D, \beta_C)$	Representation selected	Readout selected
Representation, weak	$(.5, -.15, 0)$	95	5
Representation, strong	$(.5, -.5, 0)$	100	0
Readout, weak	$(.5, 0, -.15)$	3	97
Readout, strong	$(.5, 0, -.5)$	0	100
No modulation	$(.5, 0, 0)$	50	50

Signal-bearing cases recover the generating restriction in 95–100% of replications; joint-coefficient RMSE ranges from .026 to .038. At a single condition, the constructions can be observationally equivalent, whereas crossing delay and context supplies diagnostic variation under these assumptions. These percentages are simulation recovery rates, not human study power estimates or evidence that fitted Channel weights localize storage or decision processes.

REFERENCES

Christopher J. Bates, George A. Alvarez, and Samuel J. Gershman. Scaling models of visual working memory to natural images. *Communications Psychology*, 2:3, 2024. doi: 10.1038/s44271-023-00048-3.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2): 65–70, 1979. doi: 10.2307/4615733.

John E. Kiat and Steven J. Luck. A population vector model of visual working memory for real-world scenes. *Journal of Experimental Psychology: General*, 155(5):1257–1281, 2026. doi: 10.1037/xge0001921.

Art B. Owen. The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2):386–411, 2007. doi: 10.1214/07-AOAS122.

Leonardo Pettini, Carsten Bogler, Christian Doeller, and John-Dylan Haynes. Synthesis and perceptual scaling of high-resolution naturalistic images using Stable Diffusion. *Behavior Research Methods*, 58:24, 2026. doi: 10.3758/s13428-025-02889-8.

Gaeun Son, Dirk B. Walther, and Michael L. Mack. Scene wheels: Measuring perception and memory of real-world scenes with a continuous stimulus space. *Behavior Research Methods*, 54(1):444–456, 2022. doi: 10.3758/s13428-021-01630-5.

Leo Westebbe, Yibiao Liang, and Erik Blaser. The accuracy and precision of memory for natural scenes: A walk in the park. *Open Mind*, 8:131–147, 2024. doi: 10.1162/opmi_a_00122.