

Natural scene segmentation dynamics reveal iterative Bayesian inference

Tridib K. Biswas^{1,✉}, Jonathan Vacher^{2,✉}, Sophie Molholm^{1,3,✉}, Pascal Mamassian^{4,5,✉}, and Ruben Coen-Cagli^{1,5,6,✉}

¹Albert Einstein College of Medicine, Dept. of Neuroscience

²Université Paris Cité, CNRS, MAP5

³Albert Einstein College of Medicine, Dept. of Pediatrics

⁴École Normale Supérieure Paris

⁵Albert Einstein College of Medicine, Dept. of Systems & Computational Biology

⁶Albert Einstein College of Medicine, Dept. of Ophthalmology

ABSTRACT: The visual system operates by segmenting visual inputs into distinct perceptual objects. Segmentation is dynamic, as revealed by the tempo of perceptual choices and neural activity in visual cortex. Dynamics for natural stimuli however, are poorly understood because natural scene segmentation is ambiguous and subjective. We measured subjective human segmentation maps for natural images using an innovative paradigm, and uncovered richer spatiotemporal dynamics than predicted by current theories of segmentation. To explain these dynamics, we introduced Iterative Bayesian Inference algorithms for segmentation that iteratively integrate visual inputs with the prior expectation that objects are spatially compact. When visual inputs were consistent with such a spatial prior, iterative inference was faster. This predicted relationship between spatial prior and inferential dynamics was evident in our data, and correctly reflected each individual participant’s spatial biases. We conclude that iterative Bayesian inference sets the tempo for a fundamental function of natural vision.

segmentation | psychophysics | Bayesian theories of vision | natural images | perceptual decision-making

Correspondence: ruben.coen-cagli@einsteinmed.edu

Introduction

Segmentation is the task of grouping visual inputs into the objects and parts that compose a scene, and separating those objects from each other¹. Performing segmentation supports varied functions of human vision: from reading these words, to navigating the environment, to understanding the relations between objects. In addition, segmentation influences basic processes in visual perception and may be disrupted in patients with neurologic or psychiatric disorders². Therefore, identifying the computational principles of human segmentation is crucial for understanding human visual perception and its neural substrates.

Although segmentation feels effortless and instantaneous in everyday vision, experiments have found signature dynamics in classical perceptual grouping tasks such as reporting whether two locations cued on a display belonged to the same curve^{3–7}, the same letter⁸, or the same object^{9–11}. In those tasks, reaction times vary substantially across the visual field, and the larger the distance between two parts of an image, the more time it takes to make perceptual grouping decisions^{3–7,9–11}.

An influential theory of segmentation posits that this correlation between reaction time and distance arises from how visual attention modulates the activity of retinotopically-organized neurons in the visual cortex. Namely, when attention is directed to a point in the visual field, neurons near the attended-to object in the retinotopic map should become progressively more active^{6,12} due to the sequential spreading of attention across those neurons. This theory, and related computational models^{6,11,13–15}, are further motivated by recurrent neural dynamics in early visual cortex^{16–18}.

It is unclear however, if current theories can explain the spatiotemporal dynamics of segmentation of natural images, for which segments cannot be unambiguously defined. To address this challenge, we adopted a recent experimental paradigm¹⁹ derived from a straightforward two-choice “same segment/different segments” task (Fig. 1). This enabled us to measure subjective spatial segmentation maps and relate reaction times to perceptual decision-making in segmentation. Our experiments revealed richer spatiotemporal dynamics than previously observed. Reaction time increased at larger distances for parts of the image perceived in the same segment, as in past reports. But, surprisingly, reaction time decreased at larger distances for parts of the image perceived in different segments.

We asked what computational principles could explain these spatiotemporal dynamics, and reasoned that inferences about image segments and features are coupled^{20–23}. This is because knowledge of the segments of an image helps improve the inference of the features (e.g. each segment is expected to have a particular appearance), and, in turn, knowledge of the features helps infer the segments (e.g. similar features are expected to belong to the same segment). Therefore, inference should proceed iteratively: starting from an approximate segmentation, the inference of features is updated according to those segments, then segments are revised according to newly inferred features, and so on until the quality of inference can no longer improve. This paradigm asserts that visual inference is probabilistic and requires integrating visual inputs with prior expectations²⁴ about features and segments of natural images. We refer to this theory as Iterative Bayesian Inference (IBI).

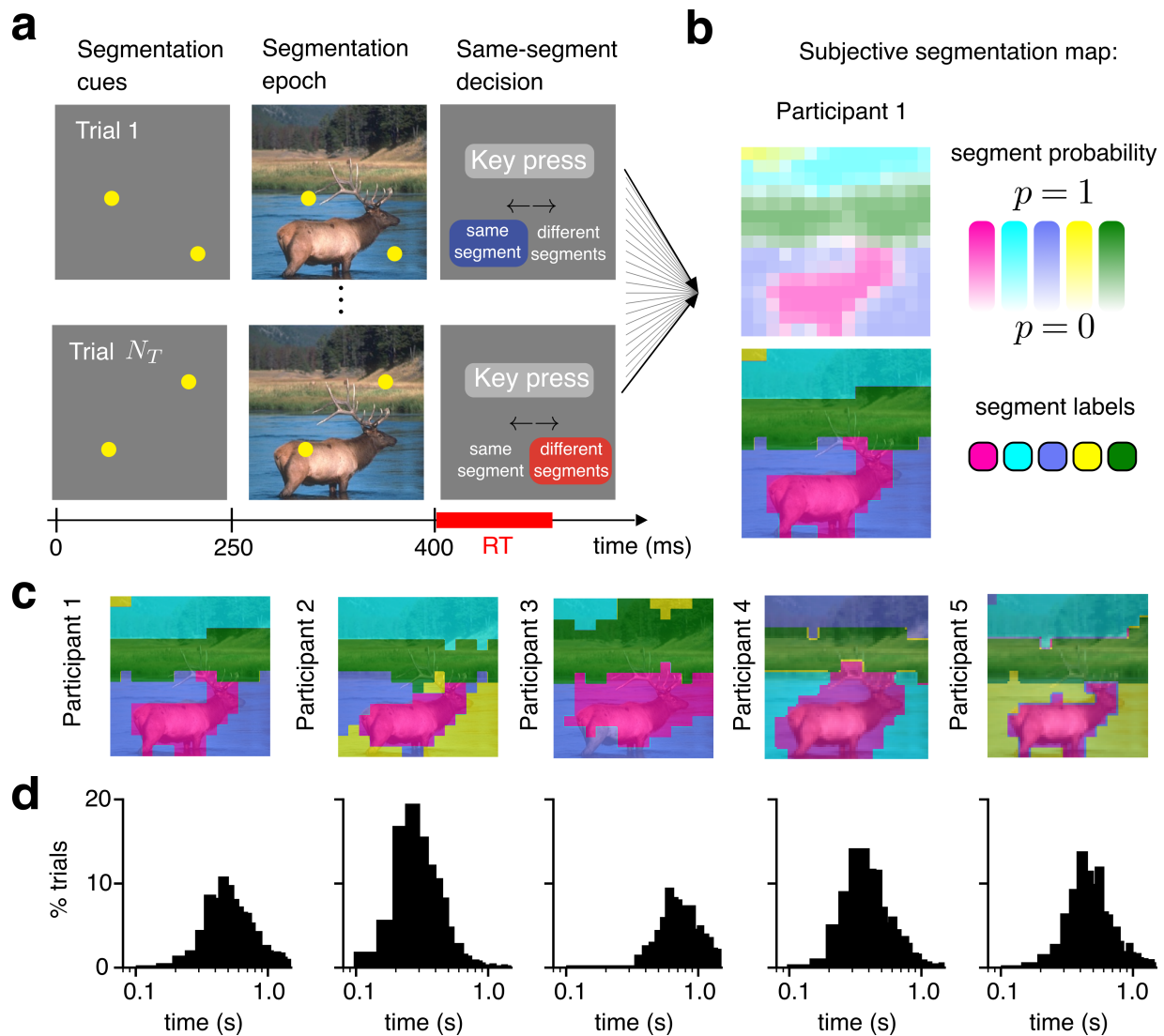


Figure 1: Experimental measurements of subjective segmentation of natural images. **a**, The participant mentally partitions the image into a pre-defined number of segments ($K = 5$ in this example) during the free-viewing epoch (not shown). Then, segmentation cues (yellow dots) are briefly shown on their own on a gray background, and subsequently on top of the image. After the offset of the image, the participant reports whether the cues were perceived in the “same segment” or “different segments”. Reaction time (RT) is measured from image offset because participants were not allowed to respond while the image was on screen. **b**, A probabilistic segmentation map (top) is decoded from participant responses over many trials (see Methods). Each color in the legend indicates a segment, while its saturation indicates segment probability. A deterministic segmentation map (bottom) is derived by assigning the most likely segment/color to each pixel. **c**, Deterministic segmentation maps are qualitatively similar across participants (colors are arbitrary and should not be compared between maps). **d**, Reaction time histograms ($N_T = 900$ trials) for all five participants in **b,c**. Bin size = 48 ms

As our theoretical analysis demonstrates, IBI theory predicts that when visual inputs are consistent with prior expectations, fewer iterations are required to improve the quality of inference. This prediction correctly captures that participant reaction times increase with distance within the same segment (because nearby parts of an image are expected *a priori* to belong to the same segment), but decrease with distance between different segments (because distant parts of an image are expected to belong to different segments). In accordance with IBI, the strength of the spatiotemporal effect we observed also reflected participant choice biases measured independently of reaction time. These empirical findings could not be explained by algorithms that either ignored the prior or substituted the iterative dynamics

of evidence with the simpler, linear dynamics used in the popular drift diffusion model^{25,26}. Our results establish IBI as a fundamental computational principle in perceptual segmentation.

Iterative algorithms like IBI motivate functional theories of neural and perceptual dynamics^{27–33} and are abundant in both classical³⁴ and modern^{13,35} machine learning. Therefore, our results have broad significance for understanding natural vision.

Results

Measuring perceptual segmentation of natural images. Our experiment adopts a novel psychophysical method to measure the segmentation of entire images¹⁹ (Fig. 1a, details in Methods). We displayed a natural image

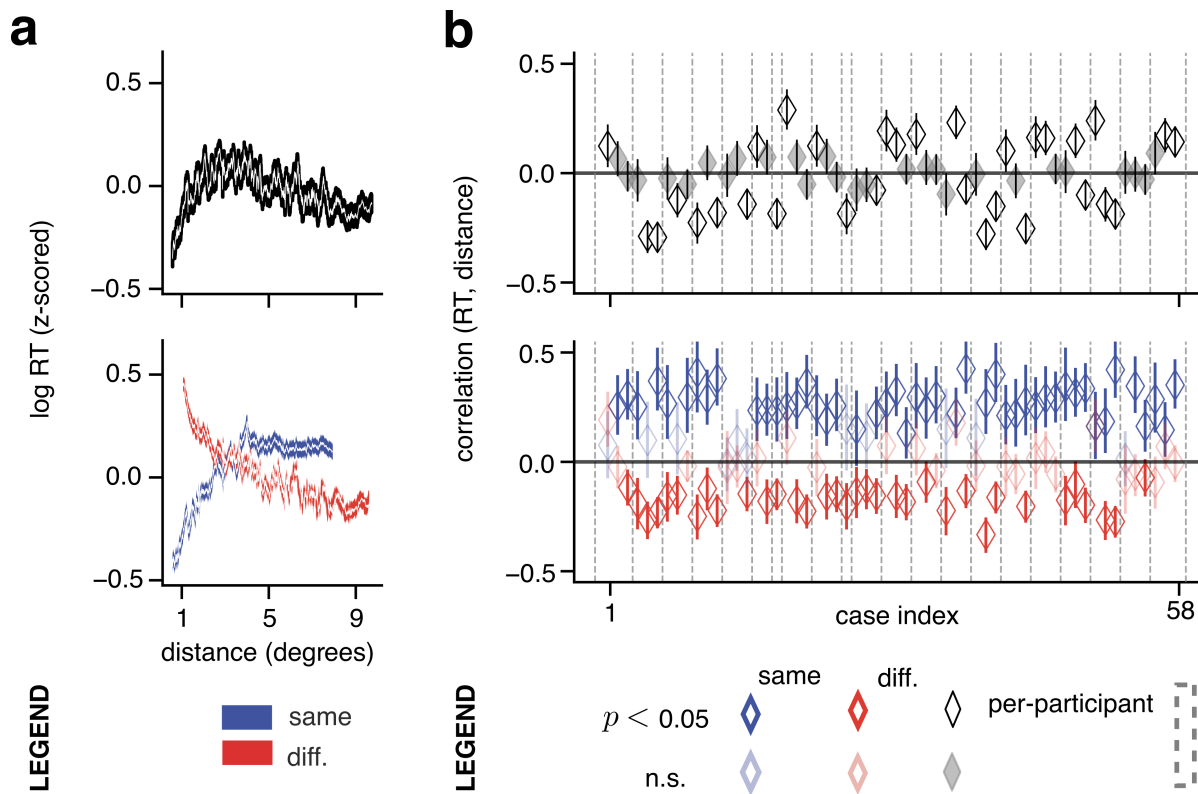


Figure 2: Opposite time-distance correlations within versus across segments. **a**, Aggregate across all cases (one case equals one image and participant). Top, z-scored log of reaction time plotted against distance (in degrees of visual angle) for all trials, sliding window kernel of 600 trials. White lines: mean. Shaded areas: s.e.m. Bottom, trials grouped by whether participants responded “same segment” or “different segments” in a given trial. **b**, Results per case. Top, black, unfilled diamonds indicate the Spearman correlation between reaction time and distance for cases with significant correlation ($p < 0.05$); gray, filled diamonds for cases with $p > 0.05$. Bottom, blue diamonds indicate the Spearman correlation for the “same segment” response subset, per case, while red diamonds indicate it for “different segments” response subset. Cases with $p > 0.05$ are plotted using faded colors. In both panels, error bars: 95% confidence interval; markers within vertical dashed boundaries indicate cases from a single participant.

to participants, and instructed them to partition it into a predetermined number of segments, in other words, to build a mental segmentation map. Then, in each trial, we presented two spatial cues in random locations on a gray background (250 ms), and subsequently presented these same cues on the natural image (150 ms). After image offset participants were asked to report with a key press, as quickly and accurately as possible, whether they perceived the two cued regions as belonging to the “same segment” or “different segments”.

From the set of responses over many trials, we decoded (see Methods) a subjective, probabilistic segmentation map for each case (*i.e.* each image for each participant, in total $N = 58$ cases across $n = 21$ participants; e.g. Fig. 1b, top). We then generated the deterministic segmentation map by assigning the most likely segment to each pixel. The segmentation maps confirmed that participants perceived meaningful visual objects as distinct segments (e.g. the elk, water, grass, and trees in Figure 1b, bottom). Maps for each image were also qualitatively consistent across participants (Fig. 1c and Supp. Fig. 1). Further visual inspection and quantification of the similarity between maps (Supp. Fig. 3) revealed fine-grained differences between individuals that captured the biases and uncertainties inherent to perceptual organization.

In addition to measuring subjective segmentation maps, we measured per-trial choice reaction times. Participant reaction time distributions were often skewed, (e.g. Fig. 1d) as is commonly observed in tasks presenting a speed-accuracy tradeoff²⁵.

In summary, our empirical measurements provide a rigorous characterization of the spatial and temporal aspects of subjective perceptual segmentation of natural images.

Spatial distance and perceived segments influence reaction times.

What is the relationship between spatial and temporal aspects of perceptual segmentation? Past work that addressed this question using both synthetic^{3,6,12} and natural stimuli^{9,11} found that reaction times increased with distances between cues. Unexpectedly, we found only a weak trend between reaction time and cue-distance across all trials (Fig. 2a, top; 35,013 total trials). In individual cases, this correlation was often not significant ($p > 0.05$ for 27/58 cases) (Fig. 2b, top, faded gray diamonds). Furthermore, among significant cases, positive and negative correlations were equally common (Fig. 2b, top, black-bordered diamonds). The mean Spearman correlation across all cases was: -0.0024 , s.e.m. ± 0.017 .

This weak time-distance correlation might appear to contradict past results. Past studies however, focused

on cues placed on the same experimentally-defined object. Therefore, we repeated the correlation analysis for only the subset of trials in which participants responded “same segment” ($n = 10,799$ total trials). This subset revealed — consistent with past findings — that reaction time tended to increase with distance (Fig. 2a, bottom, blue line). In all individual cases with significant correlation ($p < 0.05$; $N = 46/58$) the correlation was positive (Fig. 2b, bottom, blue diamonds). Across all cases, the mean correlation was: 0.23, s.e.m. ± 0.013 .

For the complementary subset of trials in which participants responded “different segments” however, ($n = 24,214$ trials), there was a decreasing relationship between reaction time and distance (Fig. 2a, bottom, red line). In all significant cases ($N = 34/58$) the correlation was negative (Fig. 2b, bottom, red diamonds). Across all cases the mean correlation was: -0.092 , s.e.m. ± 0.014 .

These findings explain why weak correlations were found when analyzing the complete set of trials. Opposite signs of correlations in the “same segment” and “different segments” subsets canceled out (noting that the “different segments” subset comprised 69% of all trials, but the magnitude of positive correlation was larger for trials in the “same segment” subset). This result suggests that existing theories of attentional spreading within an object paint an incomplete picture of the spatiotemporal dynamics of segmentation.

A normative theory of segmentation dynamics. We asked if our experimental observations could be explained by a normative theory of segmentation. The normative approach makes assumptions about the computations required to solve the segmentation problem, and then tests if algorithms that implement those computations reproduce human spatiotemporal dynamics.

Segmenting an image requires estimating the probability that each pixel belongs to any possible segment. This kind of probabilistic clustering problem is often solved with a well-established machine learning scheme that iterates between estimating the segments and estimating the visual features that comprise each segment, until convergence³⁶. We hypothesized that the visual system uses a similar probabilistic clustering strategy, with a Bayesian prior that favors spatially-compact segments, which we termed IBI.

To test this theory concretely, we adopted segmentation algorithms that implement IBI³⁷ (detailed in Methods). The algorithms use latent variables that describe how the features of an input image could be organized into segments with corresponding labels and probabilities. Figure 3a depicts these variables as nodes of a graph whose edges represent probabilistic dependencies between variables. Edges that connect neighboring nodes (Fig. 3a, thick brown lines) encourage pixels in the same neighborhood to share a segment label, and therefore define the Bayesian spatial prior.

Given an input image, the algorithm starts from an

initial guess for segment probability nodes and proceeds until a best guess is obtained at convergence (Fig. 3b). Dynamic segment probabilities, captured by eqs. 1 and 2 (see Methods), are therefore assigned to each node. Nodes are continually influenced by their neighbors leading to spatially heterogeneous computations that drive perceptual dynamics. For example, by inspection of Fig. 3b we can see that pixels at the center of the elk’s body (a spatially compact segment region) reached their final probability value faster than pixels at the elk’s horns (a spatially distributed segment region). This illustrates the distinctive property of IBI theory, that fewer iterations are needed to improve the quality of the inference when the visual input is consistent with the spatial prior³⁸ (mathematical details in Supp. Note 11). Therefore, without fitting to data, IBI makes a qualitative prediction that captures our key experimental finding; namely, the dynamics of inference are faster for neighboring pixels that belong to the same segment, and also faster for far-removed pixels that belong to different segments, because those configurations are consistent with the IBI’s spatial prior.

Iterative Bayesian Inference explains human perceptual segmentation dynamics. To test if our IBI theory captures the experimental data, we mapped the dynamic segment probabilities onto a “same segment” or “different segments” choice for our experimental task. The log-probability ratio between those probabilities (termed log-odds) represents the evidence in favor of either alternative (Fig. 3c; Methods eqs. 3,4), which is updated across iterations (Fig. 3d). The model returns a reaction time either when the evidence stops changing or reaches a predetermined boundary corresponding to high certainty that one choice is better than its alternative (Fig. 3e). To fit this model to human reaction times we introduced just two parameters (Fig. 3e, green text) and initialized the model’s segmentation based on participants’ segmentation maps (see Methods).

Within the model, the dynamics of evidence and reaction times vary widely across pairs of cues. We observed that at the first iteration, due to the initial guess, evidence in favor of “same segment” choices was generally stronger for close pairs, and evidence in favor of “different segment” choices was stronger for far pairs. As the iterations progressed, the evidence for the two conditions that were consistent with the prior led to faster reaction times (solid blue and hatched red histograms in Fig. 3e) than for the other two conditions (hatched blue and solid red histograms in Fig. 3e).

We quantified this effect across cases. Similar to the human data, there was an increasing trend for the IBI model in the subset of trials in which the model responded “same segment” (Fig. 3f, left, blue line). Conversely, for the subset of trials in which the model responded “different segments” there was a decreasing trend (Fig. 3f, left, red line). The average correlation across all cases was 0.13, s.e.m. ± 0.017 for the “same

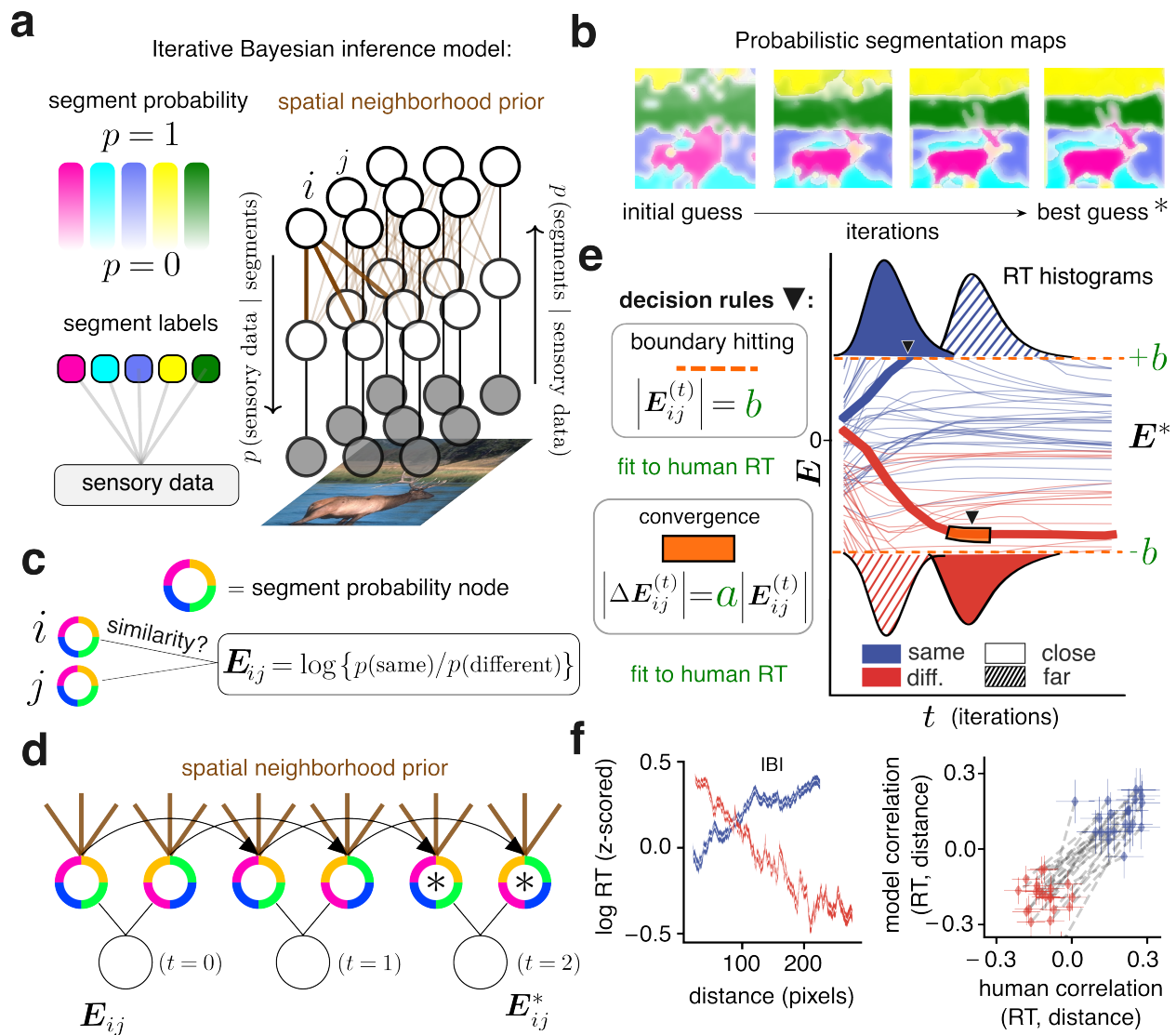


Figure 3: The IBI model predicts opposite time–distance correlations. **a**, Left: In IBI theory, sensory data extracted from the pixels of an image (bottom) is probabilistically assigned to a segment label (middle) with a corresponding probability (top). Colors are not semantic or ordered, as in Fig. 1b. Right: Inference of the segment assignment uses a probabilistic graphical model, with nodes indicating random variables and edges indicating probabilistic relationships. Brown edges indicate the probabilistic spatial prior. Arrows indicate information passed on the graph during two steps in which labels are inferred from the image features and segment probabilities of the previous step, and then probabilities and features are updated based on labels until convergence (see Methods section “Iterative inference” and Supp. 11D). **b**, Probabilistic segmentation maps, following the legend in **a**. The model starts from an initial guess for segment labels, which is iteratively refined to a best guess (*) given features in the input image. **c**, The similarity between segment probabilities of two pixels is quantified by the log–probability ratio which defines the evidence E . **d**, As the model iterates (black triangular arrows), segment probabilities are updated until the best guess (*). Evidence is computed by comparing these probabilities, which are influenced by probabilistic relationships in the spatial neighborhood prior (brown lines) at each step. **e**, Thin blue and red lines illustrate the dynamics of evidence for “same segment” and “different segments” pairs of pixels. The thick lines exemplify how a reaction time is recorded (marked by downward black triangles) when the evidence reaches a threshold value (orange dashed line) or stops increasing (orange highlights, black borders). To relate model reaction times to human reaction times two parameters (green a and b) are fit (see Methods). Top and bottom: mock RT histograms; close pairs = solid, far pairs = hatched. **f**, Left: Same conventions as in Fig. 2a, bottom, but for the IBI model, aggregated across all cases. Right: Blue diamonds indicate the time–distance correlation for the “same segment” responses subset for each participant across all images they were presented with. Red diamonds indicate the time–distance correlation for the “different segments” response subset. Gray dashed lines connect the diamonds for subsets that belong to the same participant. Error bars: 95% confidence interval.

segment” subset versus -0.2 , s.e.m. ± 0.015 for the “different segments” subset. When combining all trials of both types, these trends largely canceled each other out (correlation: -0.04 , s.e.m. ± 0.005 , across cases). Aggregated for each participant, the difference in correlations between the “same segment” and “different segments” subsets aligned with human values (Fig. 3f, right).

In summary, IBI theory reproduces the characteristic dynamics of human perceptual segmentation of natural images. It also provides a normative explanation

for why these dynamics emerge from the computation of image segments.

Stochastic linear dynamics cannot capture human perceptual segmentation dynamics.

To determine which aspects of the IBI algorithm accounted for the dynamics of human data, we compared IBI as described above to the widely adopted drift–diffusion model (DDM)²⁶ of perceptual decision making. Mathematically, the DDM has stochastic linear dynamics within each trial²⁵, in

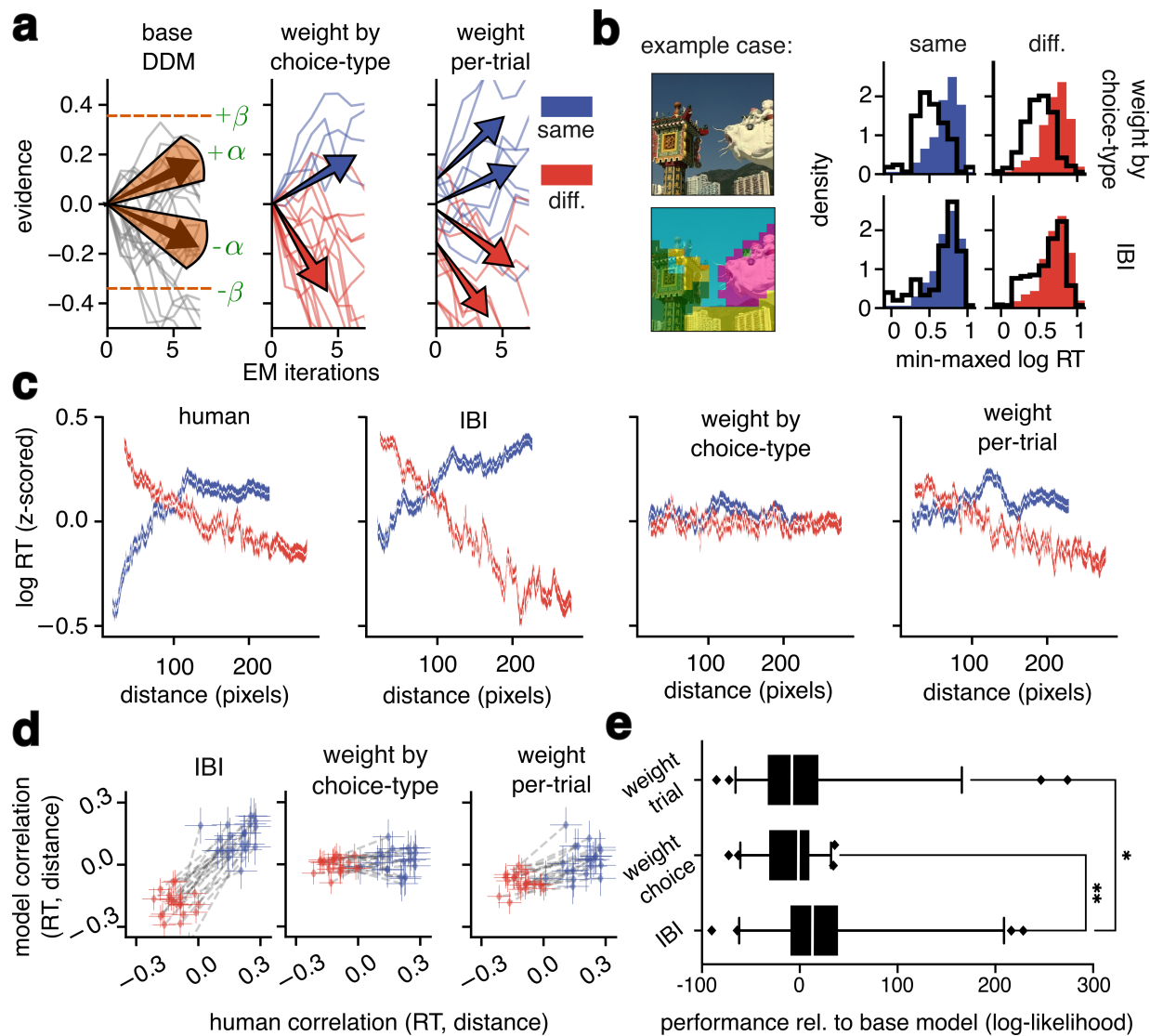


Figure 4: Drift-diffusion models do not fully capture spatiotemporal dynamics. **a**, Evidence traces from the base model (left), a choice-weighted DDM model (middle), and a trial-weighted DDM model (right); DDMs are as described in the main text. Traces are colored by the sign of the trace at the final iteration (not shown) for visual clarity. **b**, Histograms display the fit for the example case from the choice-weighted DDM (top row) vs. the fit from the IBI model (bottom row). Human reaction time distributions are filled in, while model reaction time distributions are overlaid as outlines, $n = 615$ trials. **c**, Leftmost panel: same as Fig. 2a, bottom; other panels: same plotting conventions, but for the data generated by the model class indicated in the panel title. **d**, Leftmost panel: same as Fig. 3f, right; other panels: same plotting conventions, but for the data generated by the model class indicated in the panel title. **e**, Quantitative model performance: Log-likelihoods of all image-computable models relative to the base model. Model losses were computed using 5-fold cross-validation. Whiskers indicate 5th and 95th percentile values, the Kruskal-Wallis test with Dunn’s test was used for pairwise comparisons. (* $p < 0.03$; ** $p < 0.003$).

contrast to the nonlinear within-trial dynamics of IBI. Conceptually, iterations in IBI are used to improve the quality of the approximate inference, whereas in a DDM iterations are used to average out the noise that corrupts the evidence.

To establish a reference point, we first constructed a base DDM (Fig. 4a, left) that used two fittable parameters like the IBI model. In contrast to IBI however, the base DDM used a single evidence rate that was constant within and across trials and agnostic to image input. This model did not fit reaction time data nearly as well as IBI (the IBI performance was on average 13 times better than base DDM; Fig. 4e). One possibility is that base DDM failure was due to participants in our experiments reacting systematically faster

for “different segments” choices (median decision time 480 ms, 95% c.i.±4.50 ms) than “same segment” choices (512 ms, 95% c.i.±5.44 ms; relative difference: 7%). To rule out this failure mode, we extended the base DDM by adding asymmetric evidence rates, deeming this a choice-weighted DDM (Fig. 4a, middle). The choice-weighted DDM resulted in a slightly faster median reaction time for “different segments” responses (relative difference: 0.95% in choice-weighted DDM vs. 3.43% in IBI), but did not improve histogram fits (Fig. 4b), capture distance dependence (Fig. 4c third column, 4d middle), or improve the fits quantitatively (IBI performance was on average 15 times better than the choice-weighted DDM; Fig. 4e).

The choice-weighted DDM uses a single evidence

rate for all trials with the same choice type, whereas the IBI model computes the evidence per trial based on the image content at the cued locations. To compare IBI and DDM on equal footing then, we defined an image-computable, trial-weighted DDM. We used the per-pair log-odds from IBI at convergence to weigh the DDM's evidence rate and starting-point per trial (Methods). Therefore, the trial-weighted DDM effectively included the same information about image feature similarity and the spatial prior as IBI, but maintained linear within-trial dynamics. The trial-weighted DDM improved the alignment with human time-distance correlations relative to the choice-weighted DDM (Fig. 4c,d right), but correlations remained weaker than in the actual data. Furthermore, improved alignment came at the cost of a worse quantitative fit to the empirical data (IBI performance was on average 21.2 times better than the trial-weighted DDM; Fig. 4e). Reduced trial-weighted DDMs that weighted only the drift rate or only the starting point did not improve alignment (Supp. Fig. 6). Finally, per case, the IBI model was a better fit than each of the other models in most cases (Supp. Fig. 7).

From these comparisons, we conclude that the non-linear within-trial dynamics of IBI allow the model to capture aspects of perceptual segmentation dynamics that are not captured by the accumulation of noisy evidence at a fixed rate as conceptualized in the DDM.

A Bayesian prior for spatial proximity modulates dynamics through two complementary mechanisms.

Having established the importance of modeling within-trial iterative dynamics, we set out to dissect the computational mechanisms through which the spatial prior influences those dynamics: first, the initial guess for the segment probabilities, which determines the starting point for dynamics; and second, graph connectivity (Fig. 5a, brown lines; u function in Methods eq. 2) which enforces the prior and consequently influences dynamics throughout the iterations.

The initial guess in IBI was based on the participants' own segmentation map, representing the notion of a mental map. This biased the model towards segments corresponding to objects perceived by the participants, which were often spatially compact. To remove this component of the spatial prior, we used a spatially random initialization (Fig. 5d, termed 'no mental map'). This reduced model was still able to perform the iterative computations required to update the segment probabilities, but more iterations were needed to reduce uncertainty and reach trial-by-trial decisions (compare Fig. 5e to Fig. 5c). When fitting this model to human reaction times, we observed a positive time-distance correlation for both "same segment" and "different segments" responses (Fig. 5i, middle) denoting a clear failure to capture human data (Fig. 5j, middle).

Next, we removed the graph-based local spatial prior (Fig. 5a, brown lines) by replacing all segment probability nodes with a single mean node (Fig. 5f;

termed 'no local connectivity'). The lack of local connectivity meant no information from spatial arrangement could slow down or accelerate the dynamics of evidence (Fig. 5h), and therefore decisions were made solely on the basis of image features. In order to successfully update segment probabilities, the no local connectivity model required the same initialization as the full IBI model. Nonetheless, the resulting segmentation maps were noisier than the full model and the human maps (Fig. 5g). When fit to human data, the time-distance correlations were not as aligned as the full model (note the higher negative and lower positive correlations compared to IBI and humans in Fig. 5i,j right). Furthermore, due to the noisy segmentation maps, evidence in the no local connectivity model was extremely sensitive to the precise location of the cues (see Methods and Supp. Note 11E).

In summary, perturbations of our graphical model show that the spatial prior affects within-trial dynamics through two mechanisms, and both are necessary to yield the best match to human data.

Subjective segmentation maps reveal human spatial proximity bias.

Although the spatial prior was necessary for IBI theory to capture human segmentation dynamics, it is possible that human participants did not rely on a spatial prior for their inference of segments, and that instead the time-distance correlation reflects other unmodeled factors. To address this concern, we required a model-agnostic, experimental measurement of the spatial prior that was independent of the reaction time patterns. Specifically, we reasoned that if participants relied on a spatial prior, their choices should display characteristic proximity biases.

A proximity bias for perceptual grouping of simple parametric stimuli is well established¹. In natural scenes however, it has not been measured before. The subjectivity of natural scene segments¹⁹ means that a proximity bias cannot be simply measured as a systematic deviation from an experimentally-controlled ground truth.

Our experimental paradigm offers a solution. We tested for the proximity bias by studying how participants' single-trial perceptual responses deviated from their own subjective segmentation map. In many trials ($n = 22,738/35,013$; 65%), participant responses were consistent with their subjective map (e.g. Fig. 6a, green), but in a substantial minority of trials ($n = 12,275/35,013$; 35%) the response was inconsistent with the map (e.g. Fig. 6a, lavender). For instance, the lavender-blue circles in Fig. 6a illustrate a trial in which two regions of the image were reported as being in the same segment, despite being labeled in different segments according to the subjective map (an inconsistent response).

We found that same-segment choices, whether consistent or not, were most frequent at short distances between cues (Fig. 6b both lines). Across all trials, the

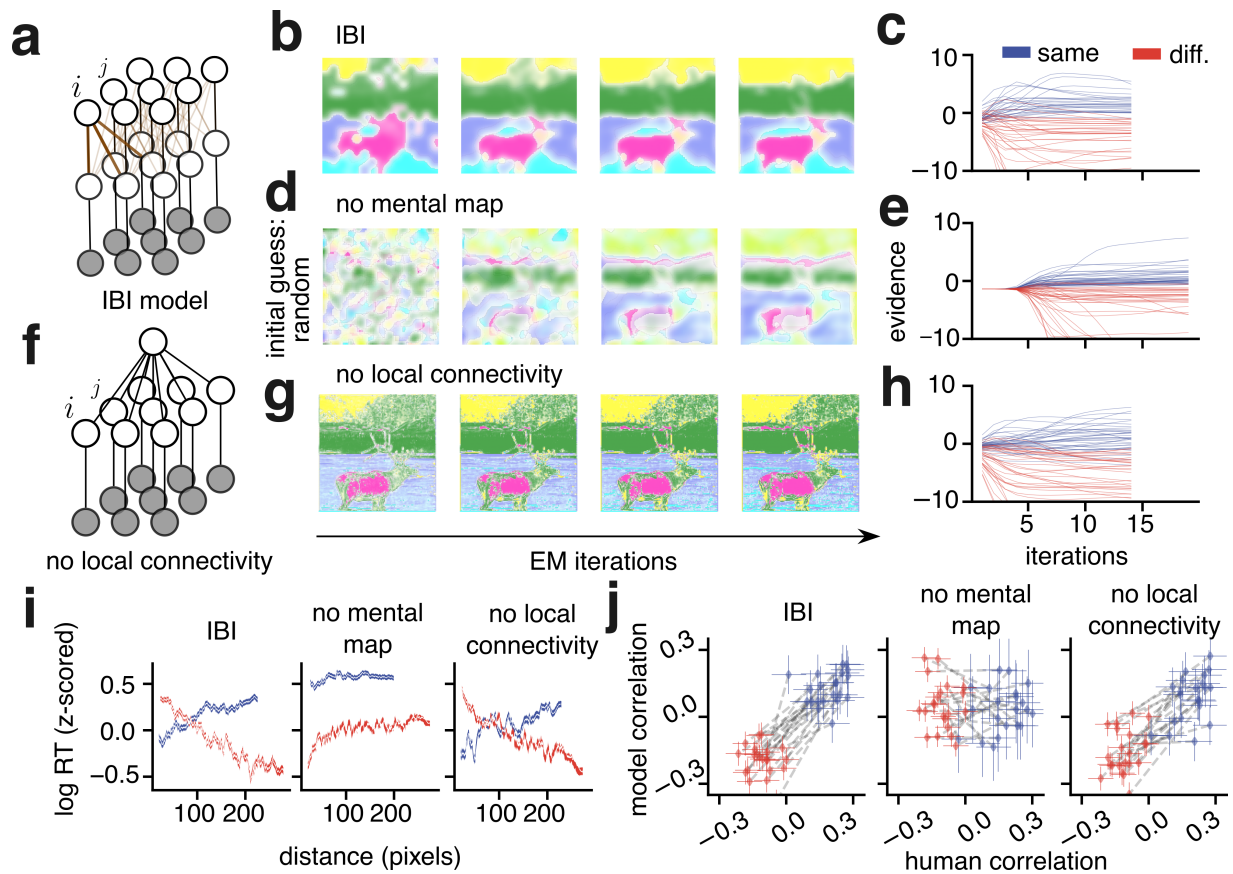


Figure 5: Algorithms that lack spatially compact segments do not capture human segmentation dynamics. **a**, The full graphical model (IBI, same as Fig. 3a). **b**, Initial iterations in the IBI model, for reference. **c**, Evidence traces from the IBI model, for reference. **d**, Initial iterations of the model when the initial guess is set to be uniformly random instead of using the participant’s segmentation map. **e**, Evidence traces that ensue from a random initialization. **f**, Perturbing local connectivity by replacing the top-layer of segment probability nodes with a single node, *i.e.* there is an identical prior probability of segments, for all pixels. **g**, Initial iterations of the model without local connectivity, but with the same initial guess as in panel **b**. **h**, Evidence traces that ensue from the no local connectivity model. Panels **b,d,g** use the same plotting conventions as Fig. 3b. Panels **c,e,h** use the same plotting conventions as Fig. 3e. Traces are colored by the choice made by the model in the corresponding trial. **i,j**, Leftmost panels: same as Fig. 3f; other panels: same plotting conventions, but for the data generated by the model class indicated in the panel title.

correlation between inconsistency rate and distance was -0.83 , s.e.m ± 0.004 ($p < 1 \times 10^{-150}$) for same-segment inconsistencies and 0.97 , s.e.m ± 0.001 ($p < 1 \times 10^{-150}$) for different segment inconsistencies. Note that, in the absence of inconsistent responses the green curve in Fig. 6b would remain at 1 at all distances and the lavender curve at 0. Instead, remarkably, the two curves were near identical, implying a very strong bias. Furthermore, if inconsistent responses were due purely to occasional attention lapses or motor errors, their proportion would be independent of the spatial separation between cues.

These results support the existence of the hypothesized spatial proximity bias for perceptual segmentation of natural images.

Human proximity bias reflects a Bayesian prior for segmentation. A general, normative prediction of our IBI theory is that when the prior is stronger, so are its effects on perceptual dynamics. This led us to two specific predictions.

First, if the measured proximity bias reflects a Bayesian prior, we should observe that the magnitude of correlation between reaction time and distance is larger for cases with larger bias. We took advantage of the

fact that, although the proximity bias was a robust feature of our data, its magnitude varied substantially. We quantified this by fitting a decaying exponential function to the proportion of “same segment” responses, as illustrated in Figure 6c ($R^2 > 0.5$ for 53/58 and 43/58 cases, respectively for trials with ground-truth “same segment” and “different segments”). Larger values of the amplitude parameter A and smaller values of the spatial decay constant κ correspond to a stronger proximity bias. We found that when the bias was stronger, the time-distance correlations had larger magnitude (correlation of bias versus the magnitude of time-distance correlation differences: 0.57 , 95% c.i. ± 0.17 ; $p < 1 \times 10^{-5}$), consistent with rational use of the Bayesian prior (Fig. 6d and Supp. Fig. 8).

The second prediction follows from the fact that the prior used by the IBI model was never fit to the human data, it was fixed in the model through local connectivity. Therefore, the difference in fit quality between models that operate with a spatial prior and models that do not, should increase for cases with larger bias. Consistent with this prediction, we found that for cases with higher proximity bias, IBI improved against the choice-

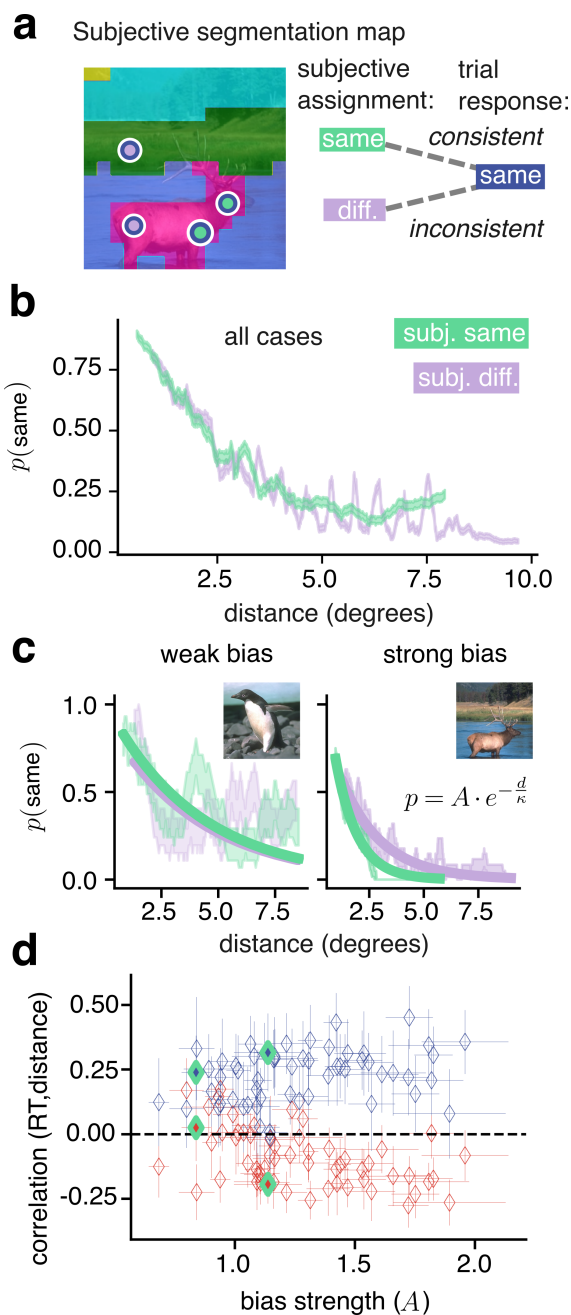


Figure 6: Consistency between responses and subjective maps reveals a spatial bias. **a**, A participant's deterministic, subjective segmentation map. Markers indicate pairs of cues presented in two trials, both with "same segment" response (blue borders). The response is consistent with the subjective map in one trial (green markers) and inconsistent in the other trial (lavender markers). **b**, Aggregate across all cases and trials, the empirical probability of responding "same segment" as a function of distance (*i.e.* the proportion of trials with response "same segment" in a sliding window of 600 trials). **c** Same as panel **b** but for two unique cases, sliding windows of 20 trials (left) and 40 trials (right). Solid lines: Exponential fit to the empirical probabilities, according to the equation in the inset. **d**, Diamonds indicate correlation of "same segment" or "different segments" subsets of trials, per case, plotted against the strength of the spatial bias at each case (*i.e.* the A best fit parameter). The left and right pairs of green-bordered diamonds indicate the left and right example cases from panel **c**. Error bars indicate 95% confidence intervals.

weighted DDM (Fig. 7a) which has no spatial prior. Furthermore, IBI also improved against the trial-weighted DDM (Fig. 7b) in which the spatial prior is only applied once to the drift and starting-point, but not applied at

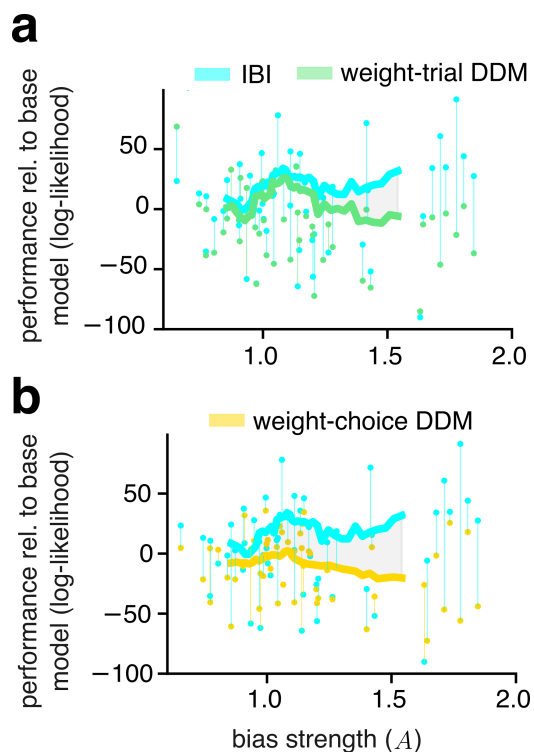


Figure 7: Cases with stronger human bias distinguish models with versus without spatial prior. **a**, Cross-validated model performance relative to the base model (defined as in Fig. 4e) as a function of bias strength. Each dot denotes model performance for one case. Each vertical line connects the performance of the IBI model (cyan) and a trial-weighted DDM (green). Thick lines: running average of the models performance with a sliding window of 15 cases. The shaded gray area indicates the difference in average performance between the two models. **b**, same as panel **a** but comparing IBI versus the choice-weighted DDM. In **b**, 6 cases with data points outside y -axis limits were excised to improve readability (but were included in the running average).

every iteration.

These results confirm that the proximity bias we measured experimentally reflects a Bayesian spatial prior for segmentation of natural images, and that the prior influences perceptual dynamics as predicted by IBI theory.

Discussion

Current understanding of perceptual segmentation and its neural basis¹² relies strongly on the classical observation that the larger the distance between two locations in an image, the longer it takes to make sameness judgments^{3-5,7,9,10}. We have shown that in natural image segmentation, this classical trend in perceptual dynamics is often inverted when the two locations belong to different segments (Fig. 2). We have proposed a theory that predicts both the classical effect and its inverse (Fig. 3), and explains the variability observed across individuals (Fig. 6,7). In our theory, dynamics emerge from iteration between inferring the features of the visual input and inferring the segments that those features belong to, with a Bayesian prior that favors spatially compact segments. Ignoring these structured dynamics of sensory inference, as is common in traditional models of perceptual decision-making, reduces the ability to

capture empirical observations (Fig. 4). Our results indicate that iterative Bayesian inference may be a fundamental computational principle for organizing natural visual inputs.

To reconcile our findings with classical experiments on segmentation dynamics, we highlight important differences in experimental design and in the underlying theoretical framework. The first experiment to report increasing reaction times with distance used a curve tracing task³. In that study, distance was defined by the length of curve between the two cues, and therefore distance could not be defined for cues on different curves. In one study that modified the task to define distance both within and across segments, the trends that emerged were similar to our findings⁴. However, motivated by the within-object focus of the sequential attention spreading hypothesis, major subsequent studies emphasized dynamics within the same curve^{5–7} or within the same object^{9,10}.

Like attention spreading, our IBI theory also centers on the dynamics of the visual representation of segments. However, in addition to considering within-object dynamics, IBI posits that the spatial prior acts across the entire visual input. This motivated us to compare dynamics within and between segments. Furthermore, as we measured subjective segmentation maps perceived by participants instead of experimentally defined segments, we tested the role of the prior in a way that previous studies could not. Namely, we have quantified the extent to which this spatial prior leads to a Gestalt-like proximity bias for each individual case. While such a bias was known for simplified and parametric stimuli¹, we have provided the first measurements in natural images (Fig. 6).

Several computational models have been developed to account for classical findings on segmentation dynamics^{6,10,11,13,14} but their applications to natural images have been limited. The ‘growth-cone’ model⁶ worked well for outline images but failed for natural images¹⁰, likely because it was not image-computable (*i.e.* it did not consider the image features). Other image-computable models were either trained only on object outlines¹³ and thus not applicable to natural images, or used hand-crafted feature detectors that are not effective for natural images¹⁴, with one recent exception¹¹. However, all of the aforementioned models produce positive correlations between reaction time and distance regardless of choice type, which is not compatible with our data.

Our IBI theory offers a probabilistic account of perceptual decision-making with natural images. According to probabilistic coding paradigms, neural representations of sensory variables are best understood as approximate Bayesian inference³⁹ that is optimized to natural input statistics^{32,40–42}. Our contributions expand the scope of those principles. First, our model is compatible with Bayesian visual-cortical representation mod-

els and substantially extends them to build richer representations of natural images: it performs joint inference of features and segments of entire images via recurrent information flow on a retinotopically organized graph. Second, our analysis reveals that the iterative dynamics required to compute joint inferences are the main determinant of the timing of perceptual segmentation decisions.

It is important to note that our results do not argue against evidence accumulation and related DDMs as fundamental in decision making. In fact, that framework provides a normative basis for how our model makes decisions using the log-odds. However, most DDM formulations are disconnected from the time-course of inferring the evidence itself. We have proposed IBI as a general framework to address this shortcoming and demonstrated that IBI models can indeed be used to define image-computable DDMs. This enables single-trial predictions for both model classes and facilitates meaningful model comparison. Model comparison demonstrated that a DDM’s ability to capture spatiotemporal dynamics of segmentation improves when it includes image information. However, we also found that within-trial inferential dynamics—which emerge naturally from IBI—are necessary to fully account for the data (although there remains room for improving the quantitative fit of IBI models to the data; Supp. 14). There are modified formulations of DDMs that capture non-linear within-trial dynamics, for instance with time-dependent decision boundaries or drift^{31,43,44}. To capture our data however, those DDMs would need to be parametrized as a function of both choice type and distance between cues. Rather than deriving new ad-hoc parameterizations of DDMs, our modeling demonstrates the advantage of considering how the visual system represents natural images. We thus show that within-trial dynamics emerge normatively from a theory that encompasses sensory inference and decision making.

Within-trial dynamics in our model also lead to predictions for neural activity. Single-neuron dynamics in V1 align well with the positive correlation between decision time and distance¹². Our novel finding, that this correlation can be reversed, raises the question of whether a corresponding signature exists in early visual cortex. In addition, correlates of the spatial prior may be evident in large-scale dynamics of spontaneous⁴⁵ and evoked¹⁵ activity, reflecting the two computational mechanisms needed to integrate the spatial prior in IBI (Fig. 5).

Although we have focused on segmentation, the insights from IBI theory are likely relevant for modeling perceptual and neural dynamics with natural stimuli more broadly given that ecological priors abound in perception^{46,47}, and that natural stimuli often engage non-trivial sensory cortical dynamics^{45,48}.

Methods

Experimental procedures.

Participants. We recruited 21 participants (11 female; age range 16-30 years) with normal or corrected to normal vision. Participants were naive to the task, and they (or their guardians, for minors) gave signed consent to participate in the experiments, and were compensated according to institutional guidelines. The study was approved by the Internal Review Board of Albert Einstein College of Medicine and Montefiore Medical Center (IRB number: 2019-10297).

Task details. We used the experimental paradigm and task design validated in our earlier work¹⁹. We first provided verbal instructions followed by a brief training session (25 trials). Each participant then completed three “blocks” (we term all trials for one image a block) separated by short breaks. At the beginning of each block, participants were presented with an image and instructed to mentally segment it with a predetermined number of segments, K , ranging from 3 to 5 (blocks were ordered pseudo-randomly). Next, the image was presented on the screen and disappeared after 5 seconds. A prompt instructed the participant to press a key to start the task. In each trial, we displayed two cues (red circles with 1-degree diameter) on a gray screen for 250 ms and then the same two cues superimposed on the image for 150 ms. Participants were not asked to maintain fixation throughout the trial but eye movements were likely minimal during the brief image presentation time. After the image and cues disappeared, participants were prompted to report whether the two cued locations belonged to the same segment or different segments. The prompt remained on the screen until participants responded with a key press. Participants were instructed to answer as quickly and accurately as possible. The following trial started after a delay of 50-250 ms.

The center of each cue was selected from coordinates defined by a square grid of size 15×15 . The grid covered the entire area of the image at lower resolution. Participants’ subjective segmentation maps were decoded at this resolution. For brevity, we refer to each pair of cues with the indices (i, j) where each i or j represents a 2D coordinate. Cues were selected pseudo-randomly with the constraint that the pair was unique in each trial. In each block, we presented the minimum number of trials required to decode the segmentation maps at the desired resolution (a small subset of all possible pairs of cues¹⁹) that depends on K : $N_T = 450, 675, 900$ trials for $K = 3, 4, 5$, respectively.

Visual stimuli and presentation. We selected 12 natural images manually from the Berkeley Segmentation Database⁴⁹, and cropped them to patches of size 256×256 pixels (see Supp. 1, 3 for image ensemble). Each image was prescribed a single value for K , and segmented by 4-6 distinct participants. Images were presented on

a gamma-calibrated LCD monitor. Participants used a chin rest to maintain a viewing distance of either 67 cm or 96 cm and we rescaled image size to span 8.7×8.7 degrees of visual angle. We combined data for the two viewing distances as no systematic differences were observed.

Data preprocessing. In each trial we recorded the binary response R_{ij} and the corresponding reaction time t_{ij} (*i.e.* the time between image offset and key press). Before analyzing reaction times, we excluded trials with t_{ij} larger than the 90th percentile (those outlier trials may correspond to attentional lapses). We verified that our main results held without this exclusion (Supp. Fig. 9), and when we controlled for sequential across-trial effects (Supp. Fig. 15). We then z -scored log transformed reaction times separately for each case (*i.e.* one image and one participant) so that the reaction time trends could be meaningfully compared across participants and images.

Decoding segmentation maps. We decoded the subjective segmentation of an image from the set of responses to all pairs presented in the experimental block as detailed in our previous work¹⁹. Specifically, we estimated the probability that any location on a 15×15 grid is assigned to any of the K segments, and termed this grid of probabilities a probabilistic segmentation map (e.g. Fig. 1b, top). This was done by numerically minimizing the squared difference between the measured response R_{ij} and the probability that the pair i, j is in the same segment, averaged over all pairs (we also confirmed that this approach did not generate signal from noise, by shuffling participant responses across pairs; Supp. Fig. 2). From the probabilistic segmentation maps, we computed a deterministic segmentation map by labeling each location on the grid with the segment that had highest probability (e.g. Fig. 1b, bottom). Lastly, we up-sampled the maps to match the resolution of the experimental image.

Iterative Bayesian Inference models.

Background. Our IBI algorithm was based on flexible probabilistic mixture models (termed FlexMM³⁷). We describe here the four main stages of the algorithm (further details are in Supplement 11).

Feature extraction. Given an input image, we first extracted features by passing the image through a deep neural network pretrained for object recognition⁵⁰. The feature space encodes more abstract information than image pixel intensity, such as edge orientation or texture. We used early-to-intermediate convolutional layers of the network because these have shown alignment with neural activity in visual cortex⁵¹. Next, we applied Principal Component Analysis to reduce the dimensionality of each layer and obtained a 12-dimensional representation \mathbf{x}_i at each pixel i (mathematical details in Supp. 11A).

Generative model. FlexMM assumes a generative model in which each observation \mathbf{x}_i depends probabilistically on a set of latent variables (*i.e.* variables inferred from observations rather than directly observed), as summarized by the graph of Fig. 3a in which nodes represent variables and edges represent probabilistic dependencies between variables³⁴. In the graph, each gray node corresponds to the image features \mathbf{x}_i for pixel i (Fig. 3a, bottom layer). The variable \mathbf{c}_i represents the segment label that the features of pixel i belong to (Fig. 3a, middle layer). The variable $\boldsymbol{\pi}_i$ (Fig. 3a, top layer) is a vector of segment probabilities, *i.e.* the probability that the features of pixel i belongs to any segment label. Given an input image, these probabilities $\boldsymbol{\pi}_i$ are what we aim to infer to obtain the probabilistic segmentation map of the image.

Following previous work relating image statistics to visual-cortical activity^{32,40–42}, we assumed that all the features \mathbf{x}_i within a given segment with label k follow a Student-t distribution with parameters collectively denoted $\boldsymbol{\theta}_k$ (see Supp. 11B for the specific parametrization). The spatial prior is enforced by introducing additional edges between $\boldsymbol{\pi}_i$ and the \mathbf{c} variables in a local neighborhood (Fig. 3a, brown edges) encouraging pixels in the same neighborhood to share the same segment label.

Iterative inference with a spatial prior. From the generative model of FlexMM, we were interested in computing the probabilistic segmentation map $\boldsymbol{\pi}$ of any given input image. Because exact computation of these probabilities is not tractable, FlexMM uses an iterative approximation scheme that extends the popular Expectation-Maximization (EM) algorithm for probabilistic clustering, to account for the spatial prior as follows. EM iterates to find the values of $\boldsymbol{\pi}$ (the organization of segments) and $\boldsymbol{\theta}$ (the organization of features within each segment) that increase the likelihood of the observations \mathbf{x} . At iteration t , the Expectation step uses the previously inferred values of the $\boldsymbol{\pi}^{(t-1)}$ and $\boldsymbol{\theta}^{(t-1)}$ to update the approximation of the posterior probability for \mathbf{c} , denoted $\boldsymbol{\gamma}^{(t)}$. Then, using the current $\boldsymbol{\gamma}^{(t)}$, the Maximization step updates $\boldsymbol{\pi}^{(t)}$ and $\boldsymbol{\theta}^{(t)}$ to their new best value, while incorporating the spatial prior by considering the current values of $\boldsymbol{\gamma}$ in a spatial neighborhood (full derivation in³⁷ and Supp. 11D):

$$\begin{aligned} \gamma_{i,k}^{(t)} &\doteq p(\mathbf{c}_i = k | \mathbf{x}; \boldsymbol{\pi}^{(t-1)}, \boldsymbol{\theta}^{(t-1)}) \\ &\propto \pi_{i,k}^{(t-1)} \cdot \text{Student}(\mathbf{x}_i; \boldsymbol{\theta}_k^{(t-1)}) \end{aligned} \quad (1)$$

$$\pi_{i,k}^{(t)} = \frac{u(\gamma_{1,k}^{(t)}, \dots, \gamma_{H \times W, k}^{(t)})}{\sum_l^K u(\gamma_{1,l}^{(t)}, \dots, \gamma_{H \times W, l}^{(t)})} \quad (2)$$

FlexMM implements the spatial prior through the function for u which must be linear but is otherwise unconstrained. We used a discrete convolution with a 2-dimensional Gaussian kernel of finite size denoted by

$H \times W$ ($H = W$ in our formulation). Following from eqs. 1 and 2, the spatial prior was introduced iteratively; the probability of the segment of each pixel was updated by averaging the previously inferred segment probabilities within a spatial neighborhood. Similarly, to further improve segmentation performance, the update of $\boldsymbol{\pi}$ at one feature layer took into account also the previously inferred segment probabilities in neighboring layers (see Supp. 11D.5 for details).

In addition, we introduced the spatial prior also through the initial guess $\boldsymbol{\pi}^{(0)}$ for each image and participant, by sampling from the participant’s subjective segmentation map which was often comprised of spatially compact segments (see Supp. 11D.6 for details).

In FlexMM, convergence of the likelihood is guaranteed³⁷ and therefore the dynamics of the probabilistic segmentation map are well defined. EM convergence is typically defined globally for the image (*i.e.* on the likelihood summed over all pixels). Yet, the segment probabilities of all pixels do not converge at the same time (mathematical details in Supplement 11D). As explained next, this is important when applying FlexMM to our experimental task.

Decision rules for IBI. Because individual pixels displayed unique dynamics, so too did pairs of pixels (see traces in Figs. 3e; 5c,e,h). Our decision rule was based on comparing the probability that a pair of pixels i, j are in the same segment (denoted π_{ij}) versus in different segments ($1 - \pi_{ij}$). At each iteration, these quantities were computed from the per-pixel probabilities and their log-ratio was used as evidence in favor of one option or the other:

$$\pi_{ij}^{(t)} = \sum_k^K \pi_{i,k}^{(t)} \pi_{j,k}^{(t)} \quad (3)$$

$$\mathbf{E}_{ij}(t) = \left[\log \left(\pi_{ij}^{(t)} \right) - \log \left(1 - \pi_{ij}^{(t)} \right) \right] \quad (4)$$

Positive values of \mathbf{E}_{ij} imply evidence in favor of “same segment” and negative values in favor of “different segments”. The model could report a decision at the first iteration for which the evidence reached a boundary b for “same segment” or $-b$ for “different segments”, which was a fittable parameter. We also included a second decision rule, to account for the fact that, for some pairs, π_{ij} might converge to a value that does not reach the boundary. Specifically, the model could report a decision when the change in evidence between consecutive iterations $|\Delta \mathbf{E}_{ij}^{(t)}|$ became smaller than the quantity $a |\mathbf{E}_{ij}^{(t)}|$, where a is the second fittable parameter. We recorded the model decision and corresponding reaction time at the first iteration in which either rule was satisfied (models that used only a single rule performed worse; Supp. Table 3, Supp. Fig. 10).

Drift Diffusion Models.

Background. We compared the within-trial dynamics of IBI against an alternative with strong support in classical decision-making literature, drift diffusion models (DDMs) ^{25,26}. We briefly review the theory before describing the bespoke, image-computable DDMs introduced in this work.

The DDM models an optimal observer—that is an observer who responds the fastest given an acceptable maximum error rate. Sequential samples of evidence, which are corrupted by independent samples of noise, are optimally integrated over time using the sequential probability ratio test, *i.e.* each new sample updates the ratio, $L(t)$, between the log-likelihood of each alternative hypothesis. DDMs assume that each sample of evidence in a given experimental condition contributes the same amount on average. Based on this assumption, the log-likelihood ratio can be parametrized in continuous time as:

$$L(t) = z + mt + \eta W(t) \quad (5)$$

Where z is a starting point, m the mean rate of evidence (drift) and $W(t)$ is a Wiener diffusion process such that: $dW \sim \mathcal{N}(0,1)$, where the variance of the diffusion process is scaled by η . In a DDM, dynamics are unique across trials because drift is perturbed by independent random draws from the Wiener process. The objective of a DDM is not to make per-trial predictions of human choices or reaction times, but to match the characteristically skewed reaction time distributions that arise from perceptual decision-making tasks ²⁵.

Base DDM. We first fit a base DDM model which was not image-computable. Defining the base DDM through equation 5, we set $z = 0$ and fit m directly to the human data per-case without any consideration of image features. η was set heuristically and fixed across cases. This procedure is in line with the typical application of DDMs in which the drift rate m is not directly encoded in stimulus information, but decoded from fitting to empirical data ²⁶.

Image-computable DDMs. Image-computable DDMs, hereafter referred to as weighted DDMs, used the segment probabilities from IBI to define the per-trial DDM parameters. In our trial-weighted DDMs we substituted m and z with the value $E_{ij}^{(t^*)}/t^*$ computed from eq. 4, obtained at time t^* which is the time of IBI’s global convergence to a final map (Fig. 4a, right; further details in Supp. 12). Reduced variants in which we substituted only one of m or z performed poorly (Supplementary Note 6). In our choice-weighted DDMs, we set $z = 0$ and simply averaged the per-trial values $E_{ij}^{(t^*)}/t^*$ over each choice type grouped by the final map at time t^* (Fig. 4a, middle).

One interpretation of using the IBI evidence at convergence to define the DDM parameters, is assuming

that the time required for sensory inference is much less than the time required for integrative decision-making subject to noise (see Supp. 12B). In other words, weighted DDMs and IBI capture the same feature and spatial information to use as evidence, but DDMs assume linear within-trial dynamics while ignoring non-linear dynamics.

Decision rules for DDM. Because DDMs assume that evidence grows at a constant rate on average, only one stopping rule is needed for when evidence hits the boundary, β . We distinguish the boundary for DDM models as β because it is fit from scratch using no information from the boundary for IBI models b . In addition, we also fit a global gain parameter α that scales the evidence for all trials.

Model fitting and comparison.

Preprocessing. Before model fitting, we normalized reaction times per case between 0 (shortest reaction time) and 1 (longest), separately for the human data and for each model.

Parameter fitting. All IBI and DDM models used two free parameters, a and b or α and β . Best fit parameters were found by maximizing the likelihood that the observed human reaction times come from the distribution of model reaction times. While some DDM’s leverage an analytical solution for this likelihood, we wanted to standardize the fitting procedure across IBI and DDMs models. Therefore, we used the empirical cumulative distribution to compute the likelihood instead. We computed such likelihoods separately for the trials in which humans chose “same segment” vs. trials in which the humans chose “different segments”, and used the negative of the sum of those two terms as the cost function to be minimized (equivalent to a maximum-likelihood approach; details in Supp. 13). This cost function was not differentiable with respect to the parameters, so we used Markov-Chain Monte-Carlo (MCMC) global minimization methods (see Supp.13 for details). Note that model parameters were not optimized to capture trial-by-trial human responses, yet model responses were largely consistent with the human subjective maps thanks to the initialization (Supp. Fig. 4).

Model comparison. For fair comparison across models we used 5-fold cross validation per case. Models were fit using trials in a training split, and then applied to predict reaction times in unseen trials that were part of a test split. All reported results are from applications to test splits *i.e.*, unseen data.

Implementation details. To add stability to all models we computed the reaction time for each pair ij over 10 pseudo-coordinates that were randomly selected to be within the spatial neighborhood of pixel i and j . Using

10 pseudo-coordinates provided 100 pseudopairs that ensured small floating-point differences from pixel to pixel did not impact model results. The code is computationally inexpensive, and can be run locally.

Statistical analysis. To plot the z -scored log of reaction time versus distance, we sorted reaction times by distance, and applied a sliding window kernel to add smoothing and elucidate trends. When comparing median reaction times for the “same segment” and “different segments” subsets, we used bootstrapping with 9999 resamples to compute 95% confidence intervals.

We quantified the relation between reaction time and distance with the Spearman correlation and we used parametric statistics to compute confidence intervals (similar results were obtained with Pearson correlation; Supp. 5).

For comparing relative likelihood distributions from each model type, we first confirmed that models were sufficiently different from each other using the non-parametric Kruskal-Wallis test. For pairwise comparisons we used Dunn’s test with a significance level of $\alpha = 0.05$; p -values were corrected for multiple comparisons.

To fit exponential curves for quantifying spatial biases (Fig. 6) we used a nonlinear least-squares regression. Parameters A and κ were optimized to fit a curve with the parametrization $p = A \cdot \exp(-d/\kappa)$, where p is the probability of responding “same segment” and d is the independent variable (distance between pixels). During fitting, we applied the bounds $[0.5, 4.3]$ (units in degrees) to κ . These were chosen as reasonable lower and upper bounds for the space constant given the size of the image. 95% confidence intervals for fit parameters were computed using two standard deviations of optimal parameter estimates.

AUTHOR CONTRIBUTIONS

T.K.B.: formal analysis, methodology, software, writing - original draft, writing - review & editing
J.V.: methodology, writing - review & editing
S.M.: resources, funding acquisition, writing - review & editing
P.M.: conceptualization, methodology, funding acquisition, writing - review & editing
R.C.C.: conceptualization, funding acquisition, methodology, writing - original draft, writing - review & editing

ACKNOWLEDGMENTS

The authors would like to thank Carlos Madrid-Aliste for assistance with running FlexMM and parameter fitting on the Einstein High-Performance Computing Cluster. We also thank Adam Kohn, Máté Lengyel, Ralf Haefner, and Odelia Schwartz for productive conversations. For participant recruitment and task administration we thank the study coordinators, Tringa Lecaj and Dennis Cregin. T.K.B. was partially supported by the National Institutes of Health Medical Scientist Training Program (T32-GM149364). R.C.C. was supported by the National Institutes of Health (EY031166; EY030578) and the Rose. F. Kennedy Intellectual and Developmental Disabilities Research Center. R.C.C. and S.M. were supported by the Simons Foundation International (SFI-AN-AR-Pilot-00009855). P.M. was supported by the French Agence Nationale de la Recherche (ANR-19-NEUC-0003-01 and ANR-17-EURE-0017). Support for recruitment of participants was provided by the Human Clinical Phenotyping Core of the NICHD funded Rose. F. Kennedy Intellectual and Developmental Disabilities Research Center (P50 HD105352).

Bibliography

1. Wagemans, J. *et al.* A century of gestalt psychology in visual perception: II. conceptual and theoretical foundations. *Psychol. Bull.* **138**, 1218–1252 (2012).

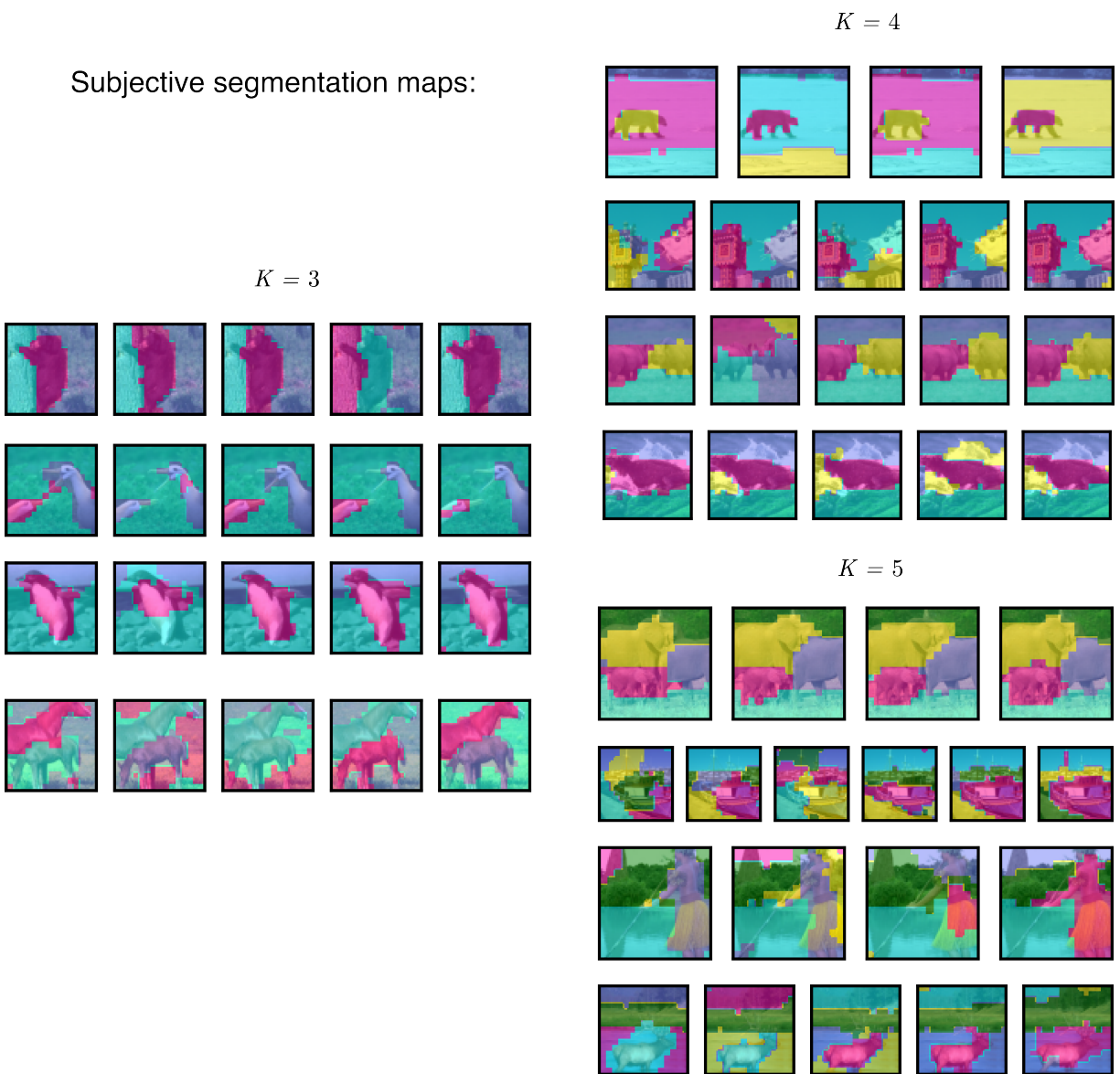
- Coen-Cagli, R. & Mamassian, P. Are we ready to tackle perceptual segmentation of natural scenes? *Vision Res.* **240**, 108749 (2025).
- Jolicoeur, P., Ullman, S. & Mackay, M. Curve tracing: a possible basic operation in the perception of spatial relations. *Mem. Cognit.* **14**, 129–140 (1986).
- Pringle, R. & Egeth, H. E. Mental curve tracing with elementary stimuli. *J. Exp. Psychol. Hum. Percept. Perform.* **14**, 716–728 (1988).
- Houtkamp, R. & Roelfsema, P. R. Parallel and serial grouping of image elements in visual perception. *J. Exp. Psychol. Hum. Percept. Perform.* **36**, 1443–1459 (2010).
- Poeresmaeili, A. & Roelfsema, P. R. A growth-cone model for the spread of object-based attention during contour grouping. *Curr. Biol.* **24**, 2869–2877 (2014).
- Ekman, M., Roelfsema, P. R. & de Lange, F. P. Object selection by automatic spreading of top-down attentional signals in V1. *J. Neurosci.* **40**, 9250–9259 (2020).
- Vecera, S. P. & Farah, M. J. Does visual attention select objects or locations? *J. Exp. Psychol. Gen.* **123**, 146–160 (1994).
- Korjoukov, I. *et al.* The time course of perceptual grouping in natural scenes. *Psychol. Sci.* **23**, 1482–1489 (2012).
- Jeurissen, D., Self, M. W. & Roelfsema, P. R. Serial grouping of 2d-image regions with object-based attention in humans. *Elite* **5** (2016).
- Adeli, H., Ahn, S., Kriegeskorte, N. & Zelinsky, G. J. Affinity-based attention in self-supervised transformers predicts dynamics of object grouping in humans. *arXiv* 2306.00294 (2023).
- Roelfsema, P. R. Solving the binding problem: Assemblies form when neurons enhance their firing rate—they don’t need to oscillate or synchronize. *Neuron* **111**, 1003–1019 (2023).
- Goetschalckx, L. *et al.* Computing a human-like reaction time metric from stable recurrent vision models. *Advances in Neural Information Processing Systems* **36**, 14338–14365 (2023).
- Marić, M. & Domijan, D. Neural dynamics of spreading attentional labels in mental contour tracing. *Neural Netw.* **119**, 113–138 (2019).
- Liboni, L. H. B. *et al.* Image segmentation with traveling waves in an exactly solvable recurrent neural network. *Proc. Natl. Acad. Sci. U. S. A.* **122**, e2321319121 (2025).
- Chen, M. *et al.* Incremental integration of global contours through interplay between visual cortical areas. *Neuron* **82**, 682–694 (2014).
- Klink, P. C., Dagnino, B., Gariel-Mathis, M.-A. & Roelfsema, P. R. Distinct feedforward and feedback effects of microstimulation in visual cortex reveal neural mechanisms of texture segregation. *Neuron* **95**, 209–220 (2017).
- Franken, T. P. & Reynolds, J. H. Grouping signals in primate visual cortex. *Neuron* **113**, 2508–2520 (2025).
- Vacher, J., Launay, C., Mamassian, P. & Coen-Cagli, R. Measuring uncertainty in human visual segmentation. *PLoS Comput. Biol.* **19**, e1011483 (2023).
- Ullman, S. Object recognition and segmentation by a fragment-based hierarchy. *Trends Cogn. Sci.* **11**, 58–64 (2007).
- Neri, P. Object segmentation controls image reconstruction from natural scenes. *PLoS Biol.* **15**, e1002611 (2017).
- Bill, J., Gershman, S. J. & Drugowitsch, J. Visual motion perception as online hierarchical inference. *Nat. Commun.* **13**, 7403 (2022).
- Shivkumar, S., DeAngelis, G. C. & Haefner, R. M. Hierarchical motion perception as causal inference. *Nat. Commun.* **16**, 3868 (2025).
- Kersten, D., Mamassian, P. & Yuille, A. Object perception as bayesian inference. *Annu. Rev. Psychol.* **55**, 271–304 (2004).
- Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J. D. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* **113**, 700–765 (2006).
- Gold, J. I. & Shadlen, M. N. The neural basis of decision making. *Annu. Rev. Neurosci.* **30**, 535–574 (2007).
- Lee, T. S. & Mumford, D. Hierarchical bayesian inference in the visual cortex. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **20**, 1434–1448 (2003).
- Friston, K. A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **360**, 815–836 (2005).
- Lengyel, M., Koblinger, A., Popović, M. & Fiser, J. On the role of time in perceptual decision making. *arXiv* 1502.03135 (2015).
- Haefner, R. M., Berkes, P. & Fiser, J. Perceptual decision-making as probabilistic inference by neural sampling. *Neuron* **90**, 649–660 (2016).
- Drugowitsch, J., Mendonça, A. G., Mainen, Z. F. & Pouget, A. Learning optimal decisions with confidence. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 24872–24880 (2019).
- Echeveste, R., Aitchison, L., Hennequin, G. & Lengyel, M. Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *Nat. Neurosci.* **23**, 1138–1149 (2020).
- van Bergen, R. S. & Kriegeskorte, N. Going in circles is the way forward: the role of recurrence in visual inference. *Curr. Opin. Neurobiol.* **65**, 176–193 (2020).
- Koller, D. & Friedman, N. *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, 2009).
- Sun, S., Li, R., Torr, P., Gu, X. & Li, S. Clip as rnn: Segment countless visual concepts without training endeavor. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 13171–13182 (2024).
- Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the *em* algorithm. *J. R. Stat. Soc. Series B Stat. Methodol.* **39**, 1–22 (1977).
- Vacher, J., Launay, C. & Coen-Cagli, R. Flexibly regularized mixture models and application to image segmentation. *Neural Netw.* **149**, 107–123 (2022).
- Salakhutdinov, R., Roweis, S. T. & Ghahramani, Z. Optimization with EM and expectation-conjugate-gradient. *ICML’03: Proceedings of the Twentieth International Conference on International Conference on Machine Learning* 672–679 (2003).
- Fiser, J., Berkes, P., Orbán, G. & Lengyel, M. Statistically optimal perception and learning: from behavior to neural representations. *Trends Cogn. Sci.* **14**, 119–130 (2010).
- Schwartz, O. & Simoncelli, E. P. Natural signal statistics and sensory gain control. *Nat. Neurosci.* **4**, 819–825 (2001).
- Coen-Cagli, R., Kohn, A. & Schwartz, O. Flexible gating of contextual influences in natural vision. *Nat. Neurosci.* **18**, 1648–1655 (2015).

42. Farzmaidi, A., Kohn, A. & Coen-Cagli, R. Relating natural image statistics to patterns of response covariability in macaque primary visual cortex. *Nat. Commun.* **16**, 6757 (2025).
43. Hanks, T. D., Mazurek, M. E., Kiani, R., Hopp, E. & Shadlen, M. N. Elapsed decision time affects the weighting of prior probability in a perceptual decision task. *J. Neurosci.* **31**, 6339–6352 (2011).
44. Deneve, S. Making decisions with unknown sensory reliability. *Front. Neurosci.* **6**, 75 (2012).
45. Berkes, P., Orbán, G., Lengyel, M. & Fiser, J. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science* **331**, 83–87 (2011).
46. Girshick, A. R., Landy, M. S. & Simoncelli, E. P. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nat. Neurosci.* **14**, 926–932 (2011).
47. Stocker, A. A. & Simoncelli, E. P. Noise characteristics and prior expectations in human visual speed perception. *Nat. Neurosci.* **9**, 578–585 (2006).
48. Kar, K., Kubilius, J., Schmidt, K., Issa, E. B. & DiCarlo, J. J. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* **22**, 974–983 (2019).
49. Martin, D. R., Fowlkes, C. C., Tal, D. & Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *Proceedings of the Eighth International Conference On Computer Vision* 416–425 (2001).
50. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations* (2015).
51. Cadena, S. A. *et al.* Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Comput. Biol.* **15**, e1006897 (2019).

Supplementary Information

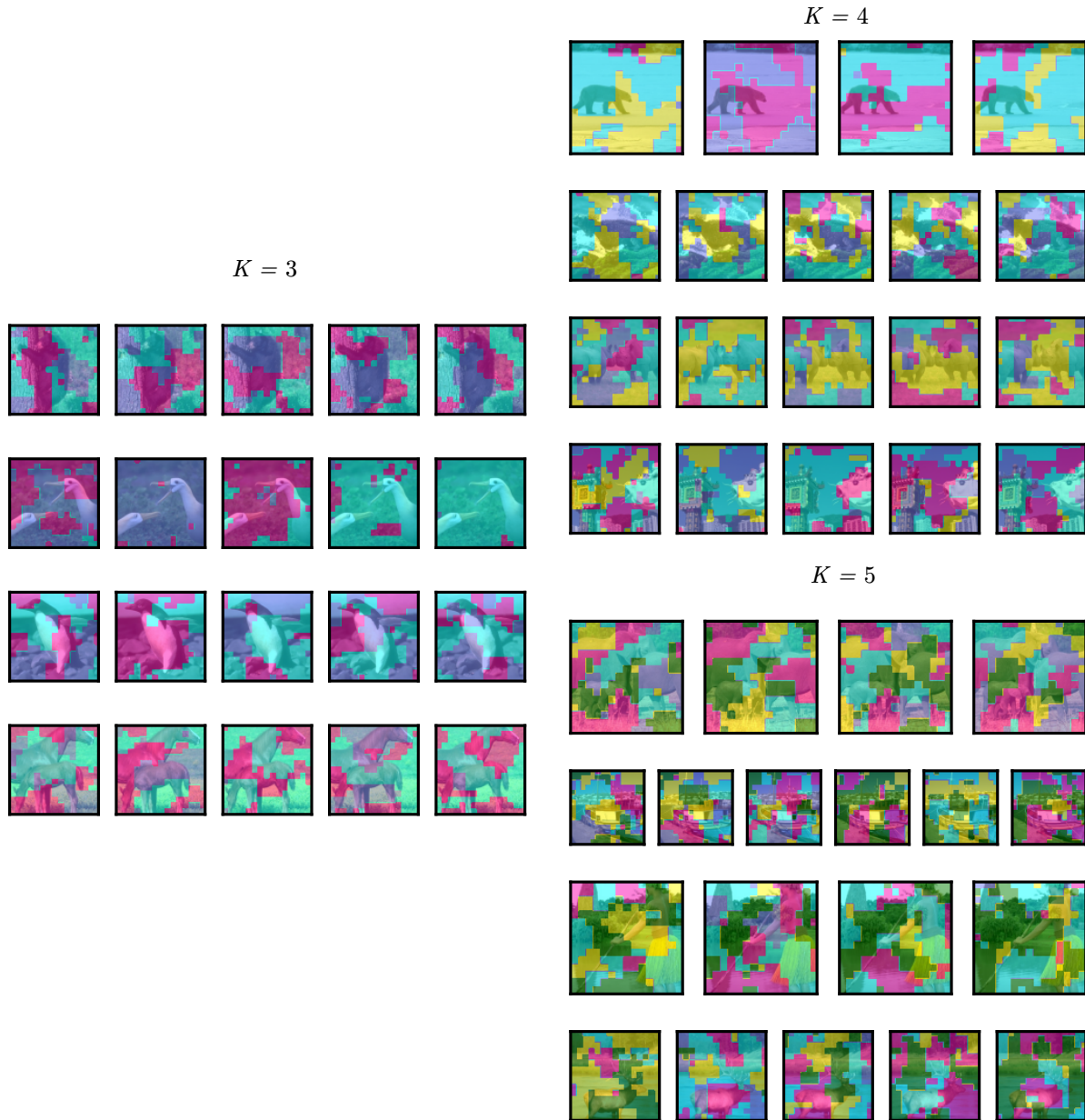
| | | |
|-----------|--|-----------|
| 1 | Subjective segmentation maps - all images | 17 |
| 2 | Control decoding of subjective segmentation maps - all images, randomized responses | 18 |
| 3 | Similarity between subjective segmentation maps - aRI comparison | 19 |
| 4 | Consistency between subjective maps and single-trial responses and between model and human responses - confusion matrices | 20 |
| 5 | Correlation between reaction times and distance – details | 21 |
| 6 | Reduced trial-weighted DDMs | 22 |
| 7 | Per-case model comparison | 23 |
| 8 | Bias against κ | 24 |
| 9 | Results without trial exclusion | 25 |
| 10 | Comparison of model performance with different decision rules | 26 |
| 11 | Mathematical details | 27 |
| A | Feature extraction | 27 |
| A.1 | Feature selection | 27 |
| B | Gaussian scale mixtures | 28 |
| B.1 | Mixtures of Gaussian Scale Mixtures | 28 |
| C | Defining the graphical model for FlexMM | 28 |
| D | Learning and inference dynamics | 29 |
| D.1 | EM as recurrent message passing | 29 |
| D.2 | EM as variational inference | 31 |
| D.3 | Per-pixel EM | 32 |
| D.4 | EM as gradient ascent | 32 |
| D.5 | Deep spatial prior | 34 |
| D.6 | EM initial guess | 34 |
| E | Graphical model perturbations | 34 |
| E.1 | No local connectivity model | 34 |
| E.2 | No mental map model | 36 |
| F | Model convergence | 36 |
| F.1 | The iterative timescale | 36 |
| G | Evidence for “same segment” | 36 |
| 12 | Drift Diffusion modeling | 37 |
| A | Sequential log-probability ratio test | 37 |
| B | Iterative and Integrative timescales | 37 |
| C | Types of DDMs | 38 |
| 13 | Parameter fitting | 40 |
| 14 | IBI model extensions | 41 |
| 15 | Reaction times decrease throughout the experimental block | 42 |

Supplementary Note 1: Subjective segmentation maps - all images



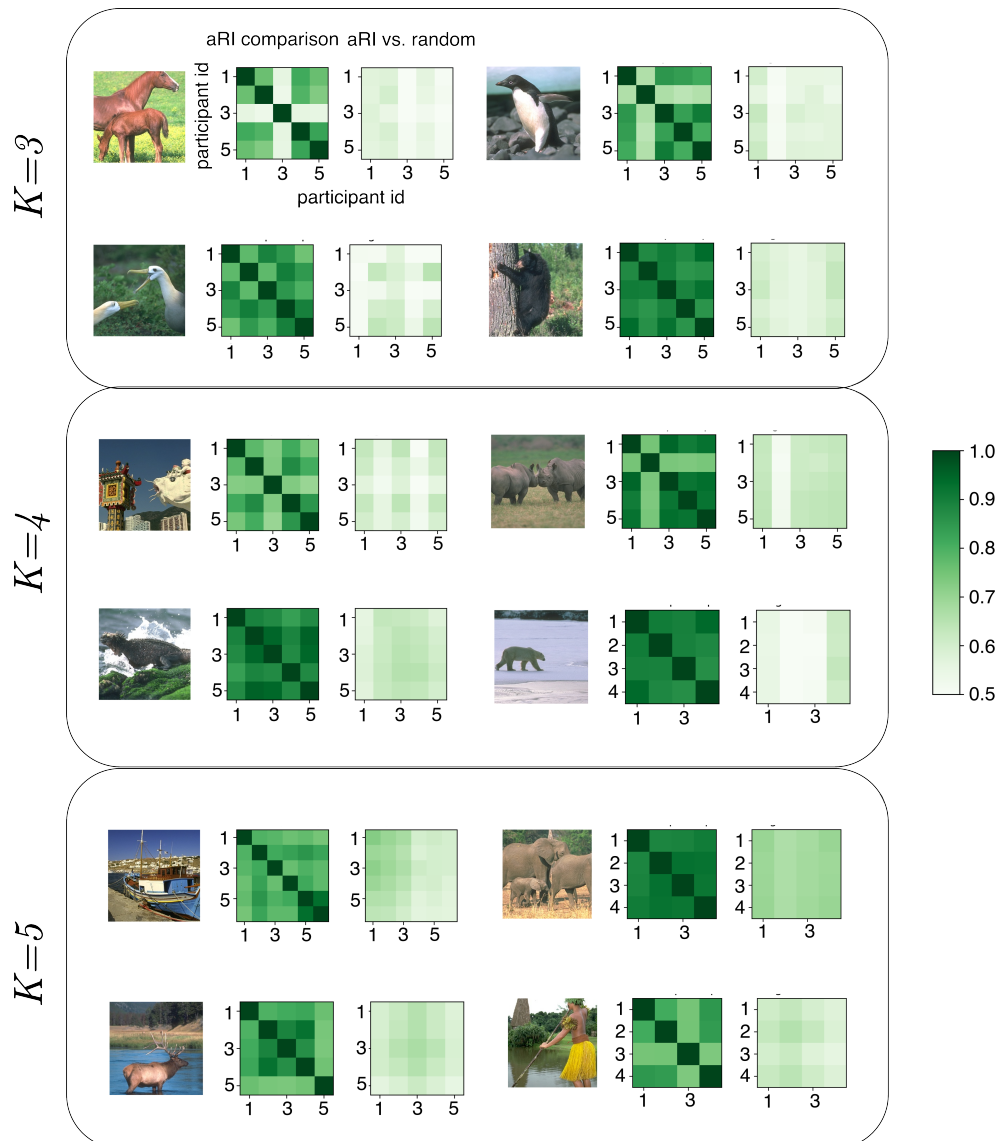
Supplementary Figure 1: Subjective segmentation maps for all images and participants. Each panel is the segmentation map for one case (one image by one participant). Images are organized by the number of segments requested, K . Each region in a 15×15 grid is assigned the most-likely segment. As in Figure 1b,c.

Supplementary Note 2: Control decoding of subjective segmentation maps - all images, randomized responses



Supplementary Figure 2: Decoded subjective segmentation maps with participant responses shuffled across trials. We devised a procedure to verify that the decoding algorithm could not hallucinate segments when the binary choice data did not contain segment-related information. To do so, we randomly shuffled the order of the responses across trials, so that each response was not aligned with the actual cue pair shown in the corresponding trial, but it was instead aligned with a different cue pair. This figure shows that, in that case, the maps produced by the decoder were not meaningfully related to the image contents.

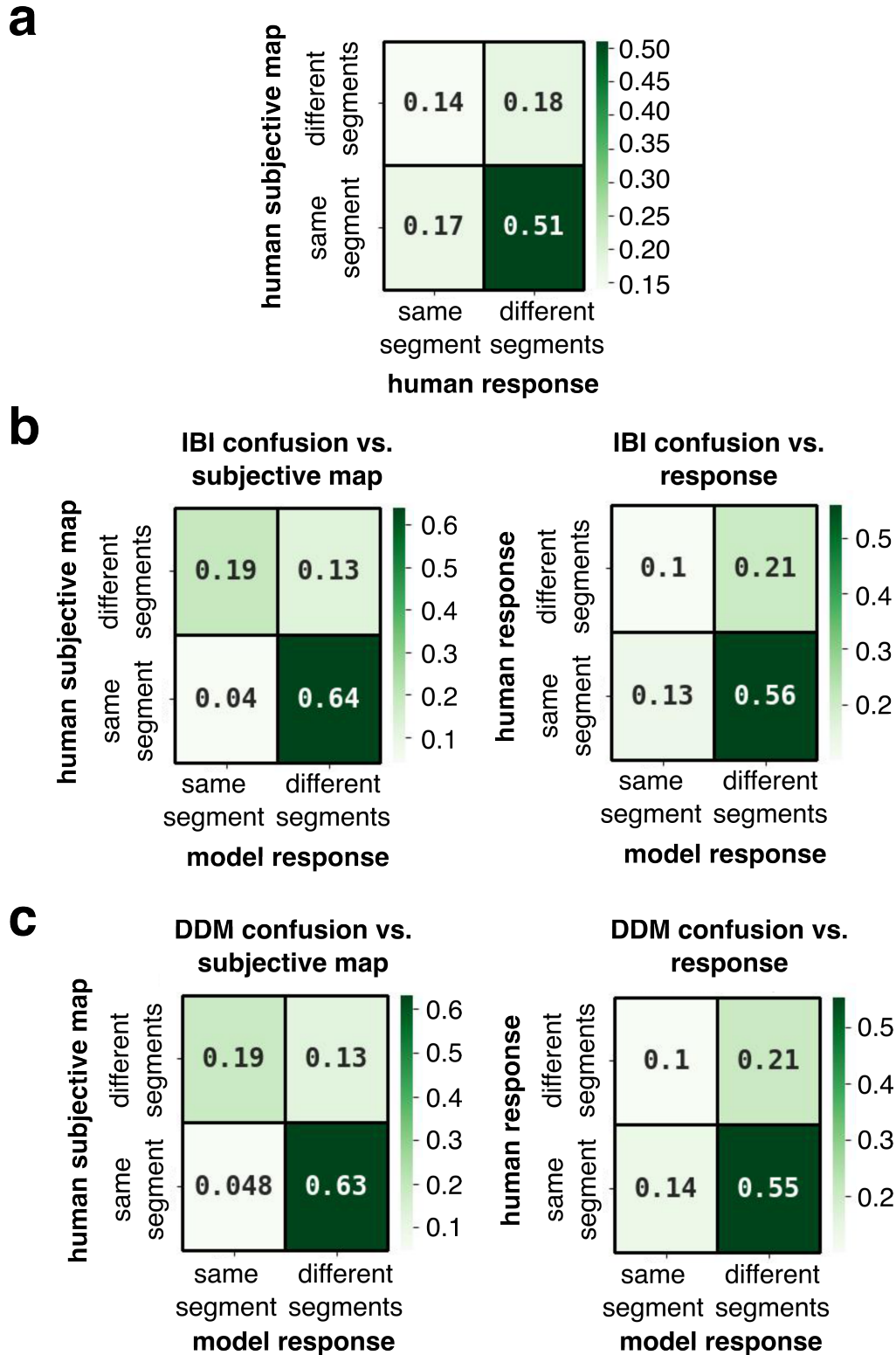
Supplementary Note 3: Similarity between subjective segmentation maps - aRI comparison



Supplementary Figure 3: Map similarity across participants. We use the adjusted Rand Index (aRI)¹ to compare the similarity of subjective maps for each image across participants. For each image, the matrix on the left represents the comparison between all pairs of participants that segmented a particular image, where each entry is the aRI comparing their maps. aRI is a symmetric measurement so the matrices are symmetric across the diagonal. The diagonal indicates a map compared to itself; by definition the value is 1. To obtain a baseline value for similarities in maps that arise from the decoding process itself, we compared the aRIs of the maps in Supp. Fig. 2 (note that the participant ids are not meaningful in the right-hand side matrix, as all participant responses are shuffled)

We use the adjusted Rand Index (aRI)¹ to quantify similarity. The aRI is a metric for comparing two clusterings that is agnostic to the cluster labels. In terms of the segmentation maps in Supp. Fig. 1, the aRI does not consider the segment colors. Instead the aRI considers all possible pairwise assignments within one map versus all possible pairwise assignments in another. A value of 1 means that all pairwise assignments are equivalent. A value of 0.5 means that, when adjusted for chance, only half of the two map's pairwise assignments agree with each other.

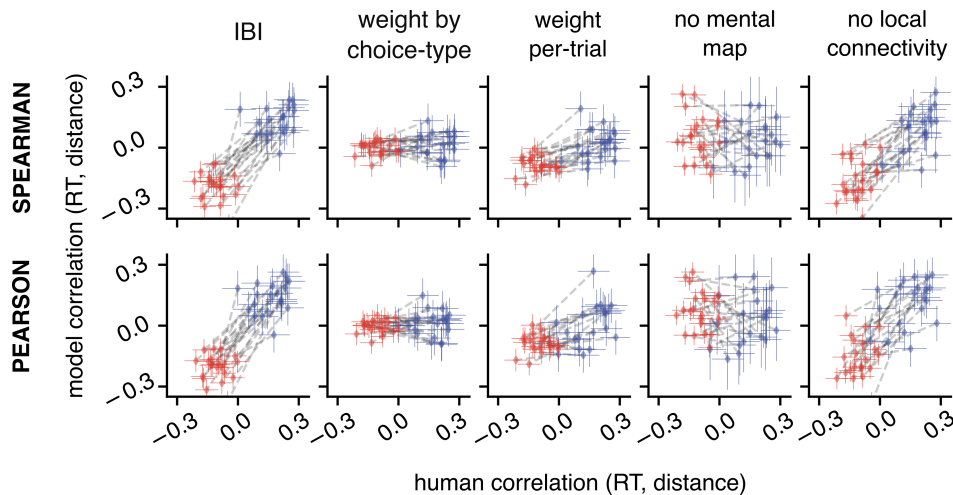
Supplementary Note 4: Consistency between subjective maps and single-trial responses and between model and human responses - confusion matrices



Supplementary Figure 4: Confusion matrices. a, Consistency *i.e.* confusion between the subjective map and single-trial responses. b, Consistency of IBI with the subjective map (left) and with single-trial responses (right). c, Here, "DDM" refers to the choice-weighted DDM. Same layout as panel b. The value reported in the confusion matrices are computed from all experimental trials, *i.e.* aggregated across cases.

Supplementary Note 5: Correlation between reaction times and distance – details

In the main text, we report results using the Spearman correlation coefficient (r_s) for two main reasons. First, the pre-existing literature makes no prediction that reaction time as a function of distance between cues should be linear, only increasing, and the Pearson correlation coefficient (r) tends to penalize non-linear increasing functions². Second, the output of our model is based on the iteration index which is an ordinal value, not a countable passing of time like human reaction times. Therefore, it is more appropriate to use rank correlation metrics such as the Spearman correlation.



Supplementary Figure 5: Comparing Spearman and Pearson correlation coefficients. Top row: Model correlations versus human correlations using the Spearman correlation coefficient (same plotting conventions as Fig. 3f, right; Fig. 4d; Fig. 5j). Bottom row: Model correlations versus human using the Pearson correlation coefficient (same plotting conventions as Fig. 3f, right; Fig. 4d; Fig. 5j)

Nevertheless, we verified that our results do not change with the use of r compared to r_s (Supp. Fig. 5). In fact, using r better differentiates the full IBI model from reduced versions in some cases (notably the no local connectivity model). To construct confidence intervals for the two metrics we used the following formulas. For the 95% confidence interval for r , where N_T is the number of trials:

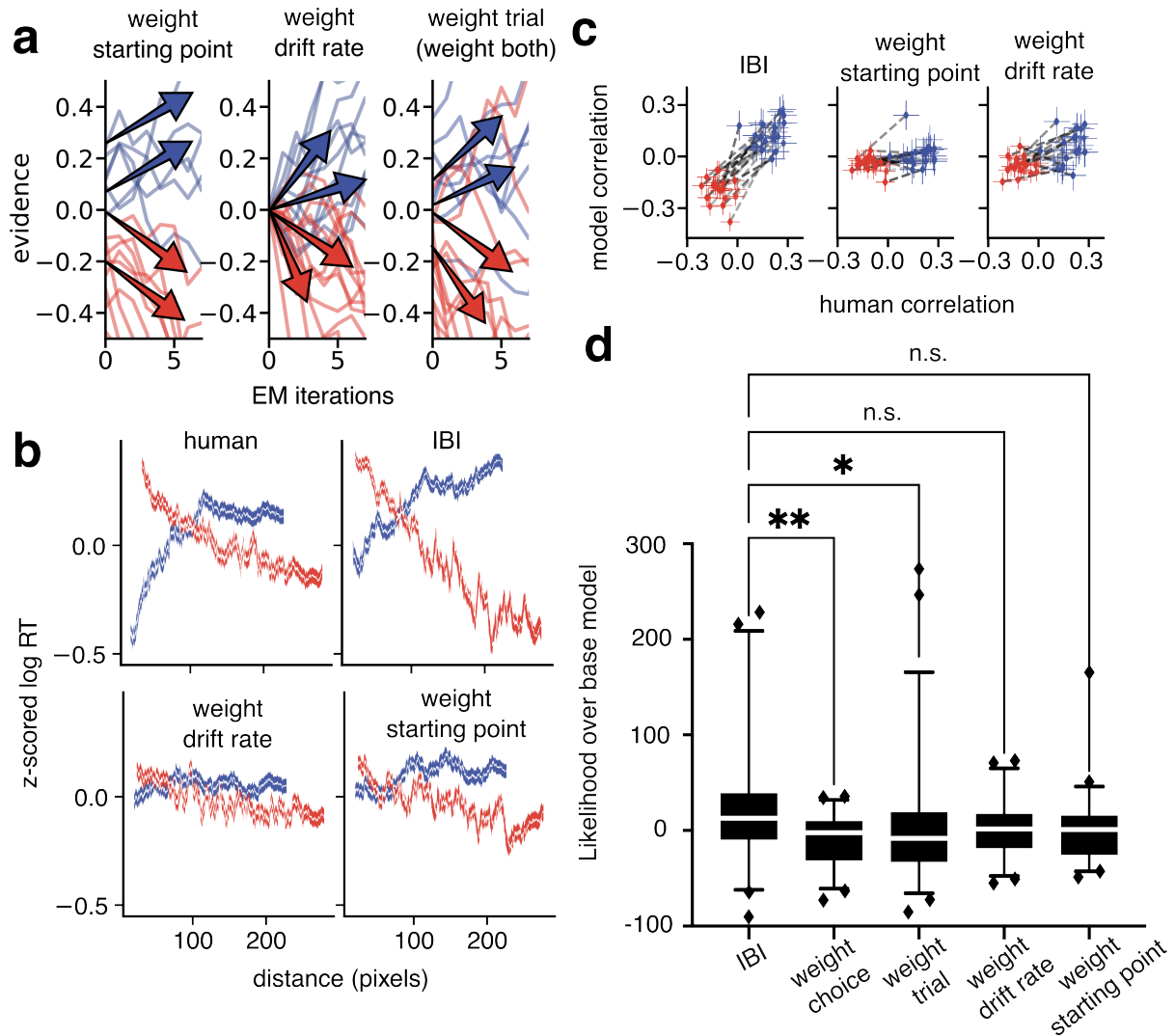
$$CI_r = \tanh \left[\arctan(r) \pm \sqrt{\frac{1}{N_T - 3}} \times 1.96 \right] \quad (\text{S1})$$

And for r_s , where $\mathcal{T}(p, N_T - 2)$ is the percentile point function for percentile p in a Student-T distribution with $N_T - 2$ degrees of freedom.

$$CI_{r_s} = \tanh \left[\arctan(r_s) \pm \sqrt{\frac{1 + r_s^2/2}{N_T - 3}} \times \mathcal{T}(0.975, N_T - 2) \right] \quad (\text{S2})$$

These equations and their justification can be found in Ruscio².

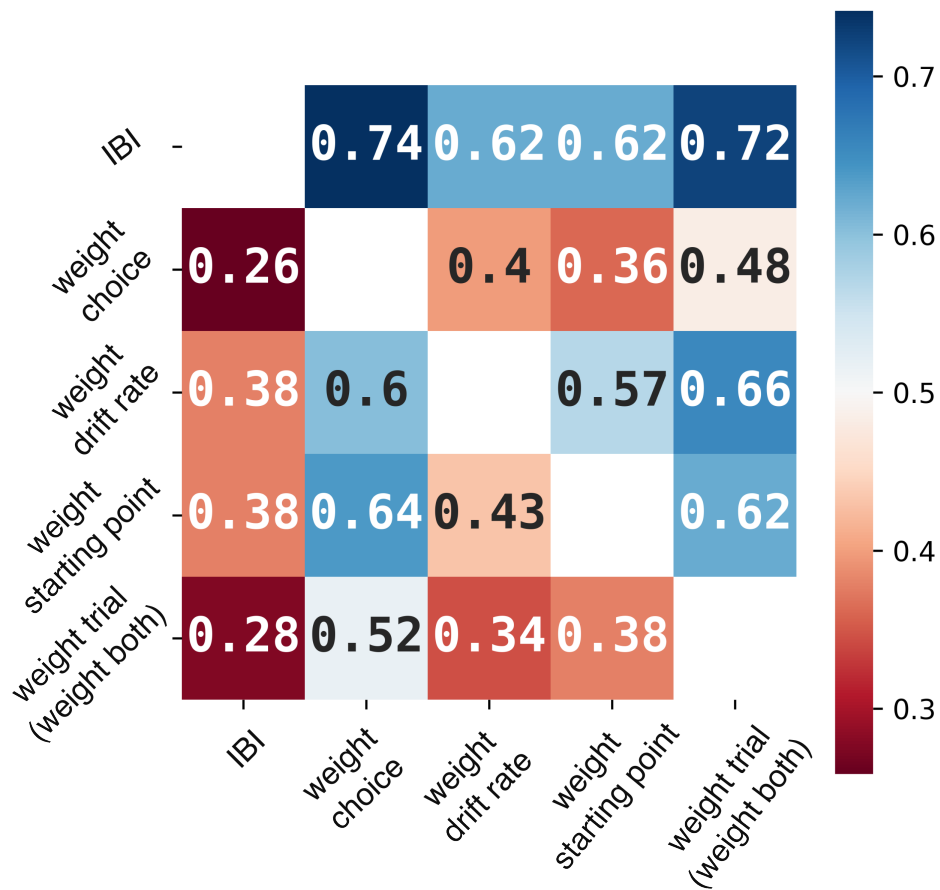
Supplementary Note 6: Reduced trial-weighted DDMs



Supplementary Figure 6: Reduced trial-weighted DDMs do not capture human correlations. **a**, Same plotting conventions as in Fig. 4a. **b**, Same plotting conventions as in Fig. 4c. **c**, Same plotting conventions as in Fig. 4d. **d**, Same plotting conventions as in Fig. 4e.

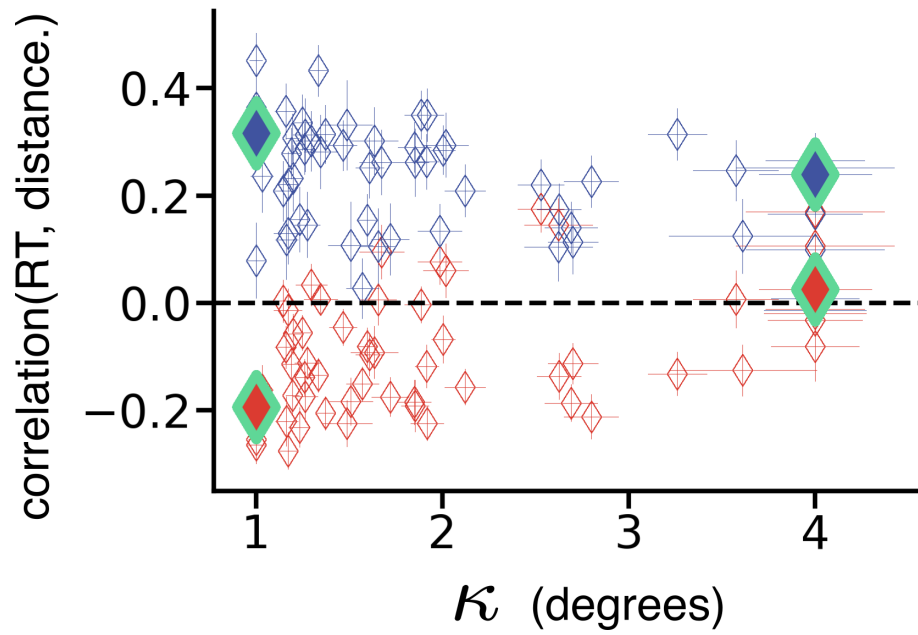
Reduced trial-weighted models are those in which only one of z or m from Methods eq. 5 is weighted. Weighing only z is referred to as the weight starting point model and weighing only m is referred to as the weight drift rate model. See Supp. 12 for precise definitions. Correlation plots in Supp. Fig. 6 are flatter than the corresponding plot for the weight per-trial model in Fig. 4d.

Supplementary Note 7: Per-case model comparison



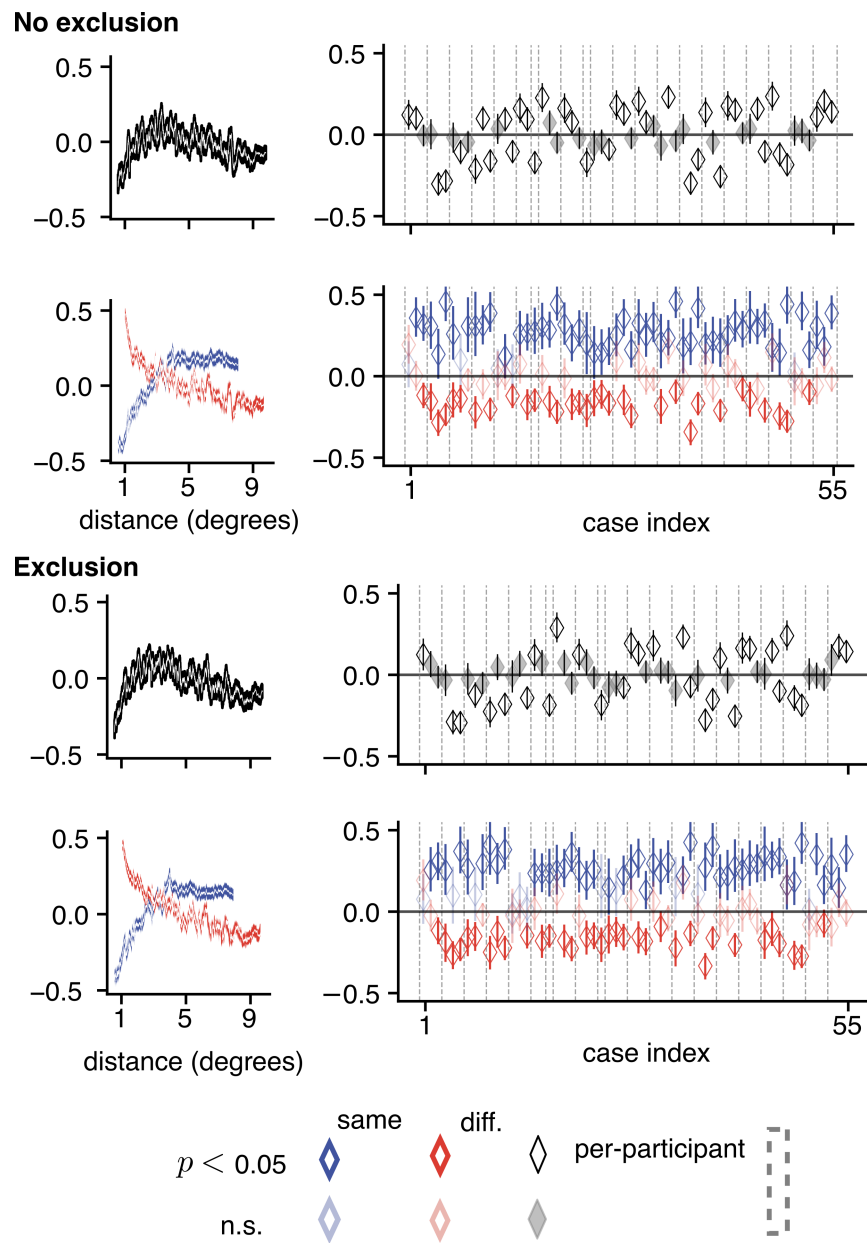
Supplementary Figure 7: Model comparison, all cases. Annotations in each cell indicate the proportion of cases in which the model specified in the row label is better than the model specified in the column label (e.g. IBI is better than choice-weighted DDM in 74% of cases, out of 58 total.)

Supplementary Note 8: Bias against κ



Supplementary Figure 8: A stronger bias leads to higher correlations. Same as Fig. 6d, as a function of the space constant of the exponential fit, κ , instead of a function of the amplitude A . Green diamonds refer to the same examples as in Fig. 6c. Note that smaller values of κ correspond to stronger spatial bias.

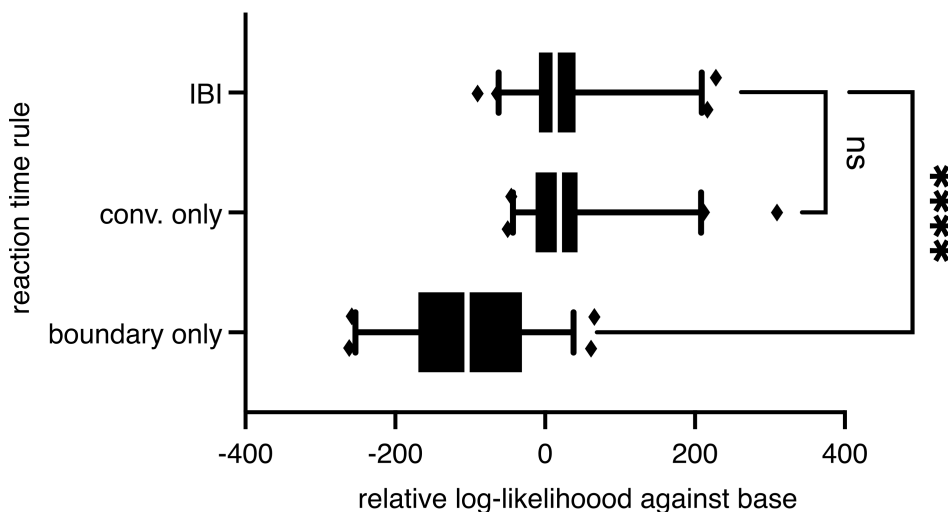
Supplementary Note 9: Results without trial exclusion



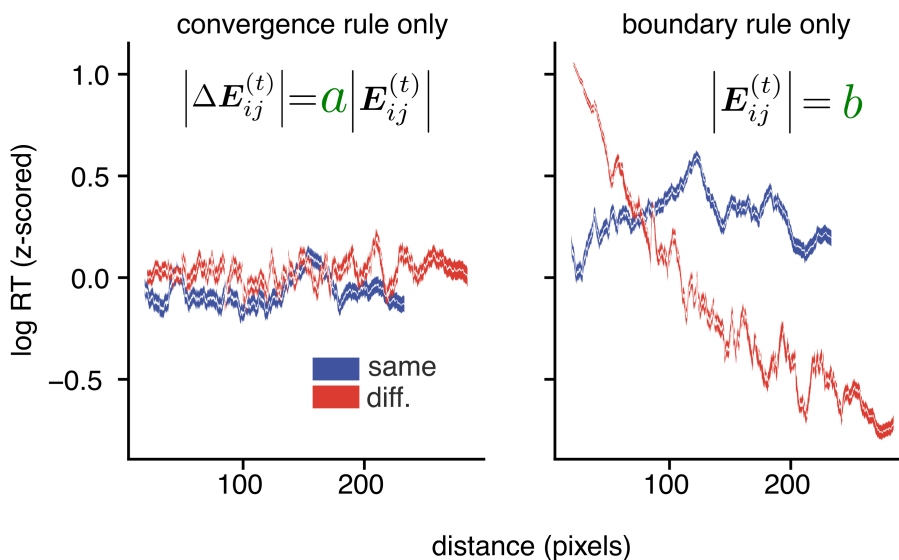
Supplementary Figure 9: Trial exclusion does not affect results. **a,c** Same as Fig. 2. **b,d** Same as Fig. 2, but without excluding the trials with the 10% slowest RT as detailed in Methods. The results are qualitatively indistinguishable.

Supplementary Note 10: Comparison of model performance with different decision rules

a



b



Supplementary Figure 10: Comparing decision rules **a**, Same plotting conventions as Fig. 4e. **b** Same plotting conventions as Fig. 3f.

To implement the convergence rule, we smoothed the evidence using a small sliding window (3 iterations) when taking the derivative. We also ensured that convergence was not transient by prescribing that the rule needed to be followed for at least 3 iterations.

The convergence rule is largely responsible for matching model distributions to human RT distributions (Fig. 10a). However, we can observe that the match to human time-distance correlations arises only when a boundary is used for the decision rule (Fig. 10b), implying that natural scene segmentation decisions involve a threshold on certainty.

Supplementary Note 11: Mathematical details

In this section, we derive the update rules presented in equations 1,2, and give an overview of the normative theory underlying our model (subsections A-D). We also discuss reduced versions of the IBI model (subsection E), and how the model produces evidence for a “same segment” or “different segments” decision (subsections F,G).

A. Feature extraction. We begin with the sensory input, an RGB image \mathbf{I} with height h and width w (in pixels). The first step is to extract features from the RGB image, which we accomplish using the deep convolutional architecture VGG-19³. Deep convolutions remap RGB values onto high-dimensional feature vectors that have been shown to be useful as predictors of neural activity in early cortex⁴. Deep neural networks have also been shown to align with the hierarchical human visual system⁵. In principle however, any shallow or deep feature extraction applied to RGB images can be used with the probabilistic inference used in proceeding steps.

Notation:

- $\mathbf{I} \in \mathbb{R}^{h \times w \times 3}$: \mathbf{I} is a tensor of shape $(h, w, 3)$ e.g. an RGB image.
- $l \in [1, 16]$: the layer of the neural network.
- $\mathbf{A}^{[l-1]}$: the activation from the previous layer of the neural network.
- $\mathbf{W}^{[l]}$: a square sliding window kernel with size 3 and feature dimension $F^{[l]}$, at the current layer. E.g. for layer 1: $F^{[1]} = 64$, $\mathbf{W}^{[1]} \in \mathbb{R}^{3 \times 3 \times 3 \times 64}$.
- $\mathbf{Z}^{[l]} \in \mathbb{R}^{h \times w \times F^{[l]}}$: the output of convolution operation
- $b^{[l]} \in \mathbb{R}^{F^{[l]}}$: the bias at a given layer

$$\mathbf{I} = \mathbf{A}^{[0]} \quad (\text{S3})$$

$$\mathbf{Z}^{[l]} = \mathbf{W}^{[l]} * \mathbf{A}^{[l-1]} + b^{[l]} \quad (\text{S4})$$

$$\mathbf{A}^{[l]} = \text{ReLU}(\mathbf{Z}^{[l]}) \quad (\text{S5})$$

The above equations describe linear-nonlinear transformations of the input as its passed through sequential layers. For all layers l , $*$ is the discrete 2D convolution operator. Values for $\mathbf{W}^{[l]}$ and $b^{[l]}$ constitute filters in an F -dimensional space, and are learned by training on the ImageNet1K database³ for object recognition (not segmentation).

For each layer $[l]$ we use a linear transformation, principal component analysis (PCA), to reduce the dimensionality of the observations⁶ to M . We set $M = 6$ to standardize feature dimensionality across layers, and because it typically explained 95% of the variance in the shallowest layers. The projection onto principal components is a simple matrix multiplication:

$$\mathbf{Z}^{[l]} \Phi^{[l]} = \mathbf{X}^{[l]} \quad (\text{S6})$$

$$\mathbf{X}^{[l]} \in \mathbb{R}^{h \times w \times 6} \quad (\text{S7})$$

where principal components are the columns of $\Phi^{[l]}$. Therefore, $\mathbf{X}^{[l]}$ is a 6-dimensional representation of the features at layer l in image data.

As in the original FlexMM implementation, to improve segmentation performance, we concatenated the low-dimensional representation at any given layer to the PCA projection of RGB features $\mathbf{X}^{[0]}$. We then defined the feature vector at layer l :

$$\vec{x}_i^{[l]} = [\mathbf{X}_i^{[0]}, \mathbf{X}_i^{[l]}] \quad (\text{S8})$$

A.1. Feature selection. For each image, we selected, out of the first five convolutional layers, the layer whose features led to model segmentation maps most similar to the participant’s segmentation maps decoded from the task (with similarity measured using the adjusted Rand Index¹). Therefore the layer of VGG-19 used was tuned to each image for each participant.

We found that intermediate layers of VGG-19 led to results best aligned with human RT-distance correlations. When features were selected from layers deeper than layer 5 or from only layer 1, model correlations were weaker. It is thought that early layers of VGG-19 have access to bottom-up image feature information such as edges, texture, color, and luminance³. Furthermore, because VGG-19 is a network trained for object recognition with backpropagation, deeper layers should have more access to object-specific or semantic information. Therefore, our feature basis partially accounts for semantic or object priors in human participants, but is not fully abstracted away from lower-level image information.

B. Gaussian scale mixtures. To simplify our notation we remove the layer-indexing from here on. Bold, italic letters indicate a vector random variable at a single pixel i ($\vec{x}_i \equiv \mathbf{x}$). Let $\mathbf{x} \in \mathbb{R}^{12}$ be the random variable representation of a single (*i.e.* at one pixel) multidimensional feature vector.

We can define the prior for \mathbf{x} rigorously through normative theories of natural vision. It has been shown that natural images present with the statistics of a Gaussian Scale Mixture (GSM), which is a standard multivariate normal $\mathbf{g} \in \mathbb{R}^{12}$ multiplied by a random global scalar $v \in \mathbb{R}^+$. The components of the multivariate normal represent the weight with which local features (such as edges with a specific orientation and spatial frequency) contribute to the input image, while the scalar multiplier represents a global modulator such as contrast^{7–11}. In our formulation, we use the Student-t distribution, which is rewritten as a Gaussian Scale Mixture by scaling the covariance of a multidimensional Gaussian, Σ , using a scalar v , that is itself a random variable sampled from a one-dimensional inverse Chi-square distribution scaled by its degrees of freedom ν . Ultimately we write:

$$\begin{aligned}\mathbf{x} &= v\mathbf{g} + \vec{\mu} \\ v^2 &\sim \nu \cdot \text{Inv-}\chi^2(\nu) \\ \mathbf{g} &\sim \mathcal{N}(0, \Sigma) \\ \mathbf{x} &\sim \mathcal{N}(\vec{\mu}, v^2 \cdot \Sigma)\end{aligned}$$

Compute the PDF of \mathbf{x} , $p(\mathbf{x})$ by marginalizing over v

$$p(\mathbf{x}) = \int_v p(\mathbf{x}|v)p(v)dv \tag{S9}$$

$$= \int_v \frac{1}{(2\pi)^6 |v^2 \Sigma|^{\frac{1}{2}}} \exp\left(\frac{(\mathbf{x} - \vec{\mu})^\top \Sigma^{-1}(\mathbf{x} - \vec{\mu})}{-2v^2}\right) p(v)dv \tag{S10}$$

The integral in equation S10 can be written as the expression f_1 ¹²:

$$f_1(\mathbf{x}, \nu, \vec{\mu}, \Sigma) = \frac{\Gamma\left(\frac{\nu+12}{2}\right) |\Sigma|^{-\frac{1}{2}}}{(\pi\nu)^6 \Gamma\left(\frac{\nu}{2}\right) \left(1 + (\mathbf{x} - \vec{\mu})^\top \Sigma^{-1}(\mathbf{x} - \vec{\mu})/\nu\right)^{\frac{\nu+12}{2}}} \tag{S11}$$

Where Γ is the gamma function. This formulation of the Student distribution makes our parametrization explicit, $\theta = (\nu, \vec{\mu}, \Sigma)$.

B.1. Mixtures of Gaussian Scale Mixtures. The objective now is to learn a *mixture* of GSMs such that each vector \mathbf{x} may be assigned to one of K classes (classes in our formulation are segments). Classes are notated as one-hot encoded vectors and we write $p(\mathbf{c})$ refer to the corresponding probability, *e.g.* if there were 3 classes, class 2 would be represented as: $[0, 1, 0]$ with corresponding probability $p(\mathbf{c}_2) \equiv p(\mathbf{c} = [0, 1, 0])$.

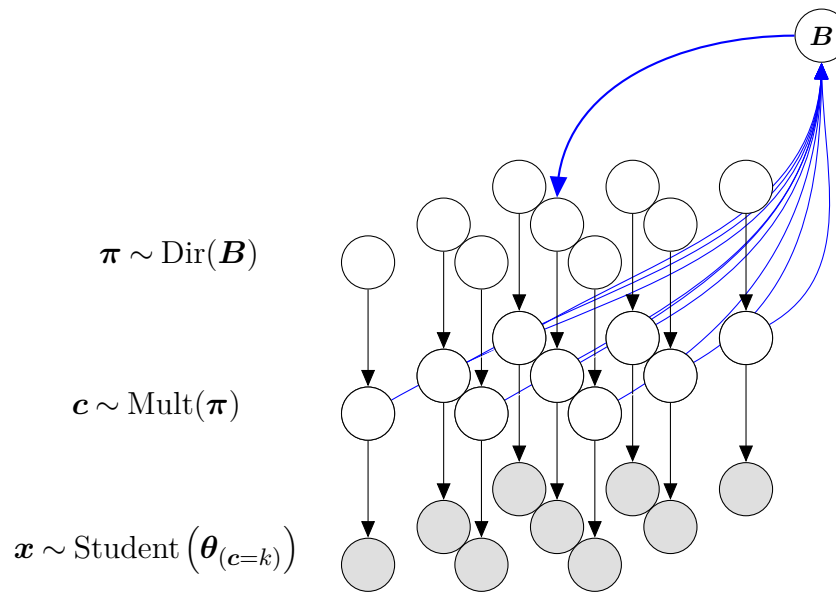
Classes are assigned with a prior probability vector denoted π which is an element of the K -dimensional simplex $\pi \in \Delta^K$, namely all its elements must sum to one. The PDF of \mathbf{x} is now simply a weighted sum:

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k f_1(\theta_k) \tag{S12}$$

Where π_k is the k th-element of the vector π , and θ_k is the set of parameters $(\nu_k, \vec{\mu}_k, \Sigma_k)$ that is unique to the k -th class.

C. Defining the graphical model for FlexMM. In Supp. Fig. 11, we define our mixture of GSMs as a graphical model, specifically a Bayesian Network, where for two random variables \mathbf{r}_1 and \mathbf{r}_2 , the notation $\mathbf{r}_1 \rightarrow \mathbf{r}_2$ indicates a conditional relationship $p(\mathbf{r}_2 | \mathbf{r}_1)$ ¹³. Following the FlexMM family of probabilistic mixture models⁶, to introduce the spatial prior, we add extra edges to the graph (Supp. Fig. 11, blue lines). These edges ultimately define a latent variable $\mathbf{B} \in \mathbb{R}^K$, the concentration parameter of the Dirichlet distribution of π . Only one \mathbf{B} node is shown in Supp. Fig. 11 for the central pixel in a 3×3 neighborhood of pixels. However, there is a \mathbf{B} node for every pixel which is influenced by that pixel's neighbors.

The extra edges inserted into the model cause a loop. Learning and inference can be performed despite the presence of the loop, as we will discuss in the next section.



Supplementary Figure 11: A probabilistic graphical model for a 3×3 pixel input. We define observed random variables as \mathbf{x} , these variables are distributed according to a Student distribution with parameters $\{\boldsymbol{\theta}\}_{c=k}$. The parameters are determined based on feature classes c . These prior probabilities of these class labels are defined by the random variable $\boldsymbol{\pi}$, distributed according to a Dirichlet distribution. A neighborhood (which is 3×3 pixels, in this illustration), of surrounding class labels defines the hyperparameter \mathbf{B} (only one \mathbf{B} is shown) which is the concentration parameter of the Dirichlet distribution. Defining \mathbf{B} on the basis of the class labels of surrounding pixels ensures that $\boldsymbol{\pi}$ variables that are spatially close to each other are more likely to be similar.

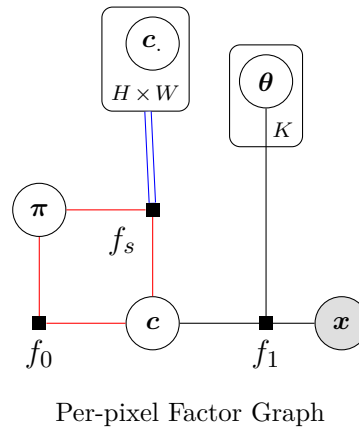
D. Learning and inference dynamics. The main goal of this section is to provide a derivation of equations 1,2 of the main text, which our results are based on. As with any probabilistic graphical model, there are multiple algorithms that can be considered for learning and inference¹³. For FlexMM⁶ we use a modified version of the Expectation-Maximization algorithm¹⁴ that guarantees closed-form linear updates for the class probabilities and flexibility in how class information is integrated across pixels to define the \mathbf{B} variables. The implementation and proof of convergence for FlexMM can be found in our earlier work⁶.

In the following sections D.1 and D.2, we focus on alternative mathematical views of our model’s dynamics in two ways, to highlight links to normative theories of perceptual dynamics. First, we briefly mention the recurrent message-passing dynamics that are present throughout the lattice of pixels to connect our work to normative theories of recurrence. Second, we use a view of EM as variational approximate inference to show that iteration is used to minimize statistical free-energy, which leads to a simplified derivation of the update rule in equations 1,2.

We note that disentangling learning (*i.e.* estimating parameters given data) from inference (*i.e.* forming the posterior distribution, given the data and parameters) requires some care for our graphical model. FlexMM alternates between an M step in which it learns parameters, including a prior probability for the class labels, and an E step in which it infers the posterior class probabilities given those parameters. In the model with the spatial prior (e.g. Fig. 3a, also Supp. Fig. 11), the parameters $\boldsymbol{\pi}$ are learned per-pixel, so the inferred posterior probability is proportional to the prior. In models without a spatial prior (e.g. Fig. 5f), a single prior probability for all pixels is learned during the M-step, and inferring a posterior probability over segments means that only the E-step generates a probabilistic segmentation map. We will discuss this distinction in greater detail in Section E when discussing the reduced versions of IBI.

In all versions of IBI and Expectation-Maximization in general, the prior is iteratively updated to generate a posterior and vice-versa. Therefore, technically speaking, the first prior our model learns is based on the initial guess which comes from participants segmentation maps. This initial guess introduces information about the segments perceived by the participants, which are likely based on both bottom-up features as well as object and scene understanding.

D.1. EM as recurrent message passing. The Bayesian Network (Supp. Fig. 11) can be rewritten per-pixel with constraints on the relationships between variables. This formalization is called a factor graph¹³. Supplementary Figure 12 is a per-pixel version of Fig. 3a and Fig. 11 formalized as a factor graph. The factor graph depicts informational relationships with undirected edges, random variable nodes \mathcal{V} and “factor nodes”, \mathcal{F} . Factor nodes \mathcal{F} elucidate the nature of the relationships between variables using the definitions of selected distributions. The joint probability distribution of the nodes \mathcal{V} is then written as the product of factors: $p(\mathcal{V}) = \frac{1}{Z} \prod_{a \in \mathcal{F}} f_a$, where Z is a normalization constant. Specifically, we use the following factorization:



Supplementary Figure 12: A factor graph of the probabilistic dependencies influencing each pixel. Compared to the probabilistic graphical model of Supp. Fig. 11, directed edges have been replaced with undirected edges and factor nodes. Rounded rectangles around nodes are referred to as “plates” and the annotation within the plate indicates how many values of that random variable are present. Specifically, there are $H \times W$ neighboring class labels used in the per-pixel class and K sets of parameters. The red lines indicate a loop in the graph. To simplify, the random variable B is not shown as a node, but is implicitly accounted for through factor f_s (eq. S15). The double blue lines indicate that information from multiple neighbors is used in the spatial prior *i.e.* c values over a local set of $H \times W$ pixels regularize the value of π . f_0 , f_1 , and f_s are factors of the joint probability distribution *i.e.* $p(x, \pi, c, c_{H \times W}, B, \theta) = \frac{1}{Z} (f_s \cdot f_0 \cdot f_1)$.

$$f_0(\pi, c) = p(c|\pi) = \text{Mult}(\pi) \quad (\text{S13})$$

$$f_1(c, x, \theta) = p(x|c, \theta) = \text{Student}(\theta) \quad (\text{S14})$$

$$f_s(\pi, c, c_0, \dots, c_{H \times W}) = p(\pi|B)p(B|c_0, \dots, c_{H \times W}) \quad (\text{S15})$$

Equation S15 describes how a spatial prior of size $H \times W$ (where $H \leq h$ and $W \leq w$, the full image size) is implemented where B is a Dirichlet parameter vector that has the same dimensionality as π ¹⁵, therefore:

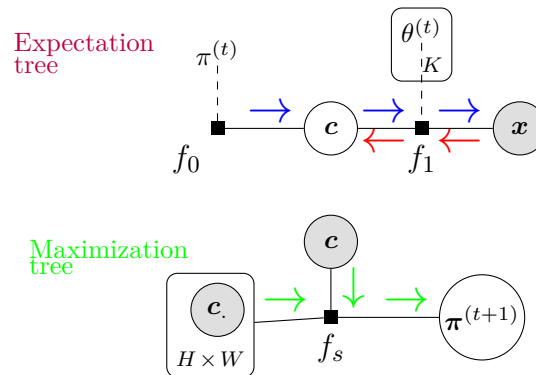
$$p(\pi|B) = \text{Dirichlet}(B) \quad (\text{S16})$$

$$p(B_k|c_0, \dots, c_{H \times W}) = \delta(u(c_{0,k}, \dots, c_{H \times W,k})) \quad (\text{S17})$$

where $u : \mathbb{R}^{H \times W \times K} \rightarrow \mathbb{R}^K$ can be any linear function, and δ is the Dirac delta function.

Our factor graph is written with a loop (Fig. 12), but previous work using the EM algorithm on loopy factor graphs¹⁶, has shown that the Expectation and Maximization steps can be considered as message-passing along two separate “tree-like” graphs, and that if neither of them contains loops the EM algorithm can be implemented exactly (see Dauwels¹⁷ for the explicit form of messages used in EM). This is the case for our model⁶ (Supp. Fig. 13).

One advantage of describing FlexMM in terms of message-passing is that it highlights the connection of FlexMM to normative and mechanistic theories that used this powerful computational scheme to clarify the role of recurrence (both lateral recurrence and feedback connections) in probabilistic neural computations^{18–20}.



Supplementary Figure 13: Breaking the loopy model using EM. The Expectation tree (top row) and the Maximization tree (bottom row) generated by “unrolling” the loopy model. Arrows represent messages passed along the edges. $\pi^{(t)}$ and $\theta^{(t)}$ are known values of parameters at the current time step and therefore not written as variable nodes here. The algorithm for computing the expectation in the Expectation Step is a forward-backward sum-product message-passing algorithm¹⁷. The blue arrows show the forward pass while the red arrows show the backward pass. In the Maximization tree, the spatial prior is implemented using max-product message passing with messages represented by green arrows¹⁷.

D.2. EM as variational inference. Having discussed the view of EM as a message-passing algorithm that connects to theories of recurrence, we now present an alternative view of EM iterations as approximate variational inference^{15,21–23}. This view emphasizes that iteration in EM is a form of minimizing statistical free-energy, and can in principle, be seen as gradient descent where information from the prior helps determine the direction of optimal descent²³. We will use this view to present an abridged derivation.

Implementing the EM algorithm as variational inference involves treating $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ as parameters in order to learn the marginal distribution $p(\boldsymbol{x}|\boldsymbol{\pi},\boldsymbol{\theta})$ that best explains the observations \boldsymbol{x} at all pixels. Here, we will not explicitly discuss the role of the Dirichlet as a spatial prior, but focus on dynamics. For a set of observations \mathbf{X} over all $h \times w$ pixels, we begin with the assumption that a single observation \boldsymbol{x}_i is independent from others. In this section, the spatial index i is explicitly written.

The log-probability of the observations factorizes as follows:

$$\log p(\mathbf{X}|\boldsymbol{\Pi},\boldsymbol{\Theta}) = \sum_i^{h \times w} \log p(\boldsymbol{x}_i | \boldsymbol{\pi}_i, \boldsymbol{\theta}) \quad (\text{S18})$$

We now introduce the latent variable \boldsymbol{c}_i , a class label, for each observation at pixel i . As in earlier sections, classes are notated as one-hot encoded vectors, *e.g.* if there were 3 classes, class 2 would be represented as: $[0, 1, 0]$ with corresponding probability $p(\boldsymbol{c}_{i,2}) \equiv p(\boldsymbol{c}_i = [0, 1, 0])$. Then:

$$\log p(\mathbf{X}|\boldsymbol{\Pi},\boldsymbol{\Theta}) = \sum_i^{h \times w} \log \sum_k^K p(\boldsymbol{x}_i, \boldsymbol{c}_{i,k} | \boldsymbol{\pi}_{i,k}, \boldsymbol{\theta}_k) \quad (\text{S19})$$

Equation S19 defines the observation likelihood as the marginal likelihood over \boldsymbol{c}_i . To move the summation outside the logarithm we can introduce an approximate distribution for \boldsymbol{c}_i , namely $\boldsymbol{c}_i \sim q^{(t)}$.

$$= \sum_i^{h \times w} \log \sum_k^K \frac{q^{(t)} \cdot p(\boldsymbol{x}_i, \boldsymbol{c}_{i,k} | \boldsymbol{\pi}_{i,k}, \boldsymbol{\theta}_k)}{q^{(t)}} \quad (\text{S20})$$

The above can be rewritten using the expectation operator $\mathbb{E}_{\boldsymbol{c}_i \sim q^{(t)}}$, which is the expectation with respect to the approximate distribution $q^{(t)}$. Using the expectation operator:

$$= \sum_i^{h \times w} \log \left(\mathbb{E}_{\boldsymbol{c}_i \sim q^{(t)}} \left[\frac{p(\boldsymbol{x}_i, \boldsymbol{c}_i | \boldsymbol{\pi}_i, \boldsymbol{\theta})}{q^{(t)}} \right] \right) \quad (\text{S21})$$

$$= \sum_i^{h \times w} \log \left(\mathbb{E}_{\boldsymbol{c}_i \sim q^{(t)}} \left[\frac{p(\boldsymbol{c}_i | \boldsymbol{x}_i, \boldsymbol{\pi}_i) \cdot p(\boldsymbol{x}_i | \boldsymbol{\theta})}{q^{(t)}} \right] \right) \quad (\text{S22})$$

Using Jensen’s inequality we can move the logarithm inside the summation, and establish a lower bound to maximize.

$$\geq \sum_i^{h \times w} \mathbb{E}_{\boldsymbol{c}_i \sim q^{(t)}} [\log p(\boldsymbol{c}_i | \boldsymbol{x}_i, \boldsymbol{\pi}_i)] + \mathbb{E}_{\boldsymbol{c}_i \sim q^{(t)}} [\log p(\boldsymbol{x}_i | \boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{c}_i \sim q^{(t)}} [\log q^{(t)}] \quad (\text{S23})$$

Inference aims to find the approximate distribution q to maximize the above lower-bound monotonically. This is functionally equivalent to minimizing the negative of this quantity, which can be interpreted as a quantity called variational free energy^{21,22,24} (note that the third term is the statistical entropy of the approximate distribution $q^{(t)}$).

Note that the challenge with performing inference with the model as written in equation S24 is that in the first term:

$$\sum_i^{h \times w} \mathbb{E}_{\boldsymbol{c}_i \sim q^{(t)}} [\log p(\boldsymbol{c}_i | \boldsymbol{x}_i, \boldsymbol{\pi}_i)] \quad (\text{S24})$$

there are $(K - 1) \cdot h \cdot w$ parameters and only $h \cdot w$ observations. Therefore, to effectively reduce the number of parameters we use regularization: we introduce a linear function u that can be applied over the $\boldsymbol{\pi}_i$ within a neighborhood of size $H \times W$ to facilitate spatial smoothing. In practice, special care must be taken to define the u as linear using additional parameters from conjugate priors (see previous Section D.1 and the FlexMM paper⁶, particularly

Appendix B for proofs). We ignore these additional parameters in the current derivation to focus on one main point — the u function must be linear as applied to the expectation so that it can be taken out of the expectation term by the linearity of expectation:

$$\mathbb{E}_{\mathbf{c}_i \sim q(t)} [u(\log p(c_{i,k} | x_i, \pi_{i,k}))] = u_{i,k} \left(\mathbb{E}_{\mathbf{c}_i \sim q(t)} [\log p(c_{i,k} | x_i, \pi_{i,k})] \right) \quad (\text{S25})$$

To apply the condition that closer pixels are more likely to be in the same class we use as the u function a convolution with a discrete 2D Gaussian with finite kernel size.

D.3. Per-pixel EM. Notice that the linear u function can be applied to the standard formulation of the expected complete-log-likelihood over neighbors. The standard formulation of EM states that parameters evolve to increase the likelihood of the dataset as a batch¹⁴. In our case, parameters are per-pixel. Examining the likelihood traces in our model we found that certain pairs of pixels reach their final likelihood value before other pairs. Therefore it is worth a brief look at existing theory on EM with single observations, termed incremental EM, which was shown to be feasible through the variational view of EM²¹. Neal and Hinton have shown that maximizing $\log p(\mathbf{X} | \mathbf{\Pi}, \boldsymbol{\theta})$ is equivalent to maximizing the complete log-likelihood per-pixel when latent observations $\{\mathbf{c}\}_{\forall i}$ are independent. In our formulation strictly speaking, $\{\mathbf{c}\}_{\forall i}$ are not independent because of the u function. However, because $\boldsymbol{\pi}_i$ are also learned per pixel we can treat $\{\mathbf{c}\}_{\forall i}$ as conditionally independent when $\boldsymbol{\pi}_i$ is known for all i . This justifies that single pixel updates are unique while increasing likelihood for the whole image.

D.4. EM as gradient ascent. Having established that we can apply u to the standard complete-log-likelihood, and that this quantity can be optimized per pixel, we can introduce a per-pixel objective functional \mathcal{O}_i by defining a parameter space $\{\boldsymbol{\pi}_i, \boldsymbol{\theta}\} \in \mathcal{V}$ and a distributional space $q \in \mathcal{Q}$. The way that iteration enables gradient ascent for this functional is summarized in Table 1, and aligns with normative work exploring dynamic sampling circuits^{25–27}. The equation for the per-pixel objective functional at time step (t) (ignoring the entropy contribution from equation S23) is:

$$\mathcal{O}_i(\mathcal{Q}, \mathcal{V}) = \mathbb{E}_{\mathbf{c}_i \sim q(t)} [\log p(\mathbf{x}_i, \mathbf{c}_i | \boldsymbol{\pi}_i, \boldsymbol{\theta})] \quad (\text{S26})$$

Consider the derivative $\frac{\partial \mathcal{O}_i}{\partial \mathcal{Q}}$. Past work has shown²¹ that when setting $\frac{\partial \mathcal{O}_i}{\partial \mathcal{Q}} = 0$ a maximum exists at: $q^\dagger = p(\mathbf{c} | \mathbf{x})$. At any given iteration t , computing $q^{\dagger(t)} = p(\mathbf{c} | \mathbf{x})$ is the expectation step, while the maximization step involves learning parameters $(\boldsymbol{\pi}_i^{(t)}, \boldsymbol{\theta}^{(t)})$ to further improve approximation. The EM update rule as derived below uses parametrized distributions to make these computations exactly, but it is nevertheless a specialized form of gradient-based optimization with the goal of approximation. These closed-form expressions meaningfully capture dynamics because constrained optimization (*i.e.* $\boldsymbol{\pi}$ must sum up to one) require a small step size. This was shown by Xu and Jordan²⁸ who explicitly relate EM to iterative Newton and Quasi-Newton’s methods.

To continue the derivation, for the M step we re-express the expectation in equation S26 explicitly as a sum over class labels and we introduce explicitly the time labels:

$$\mathcal{O}_i(q^{\dagger(t)}, \mathcal{V}) = \sum_k^K (\log p(\mathbf{x}_i, c_{i,k} | \pi_{i,k}, \boldsymbol{\theta}_k)) \cdot p(c_{i,k} | \mathbf{x}_i, \pi_{i,k}^{(t-1)}, \boldsymbol{\theta}_k^{(t-1)}) \quad (\text{S27})$$

We use Bayes’ rule for the conditional probability term:

$$p(c_{i,k} | \mathbf{x}_i, \pi_{i,k}^{(t-1)}, \boldsymbol{\theta}_k^{(t-1)}) = \frac{p(\mathbf{x}_i | c_{i,k}, \boldsymbol{\theta}_k^{(t-1)}) p(c_{i,k} | \pi_{i,k}^{(t-1)})}{p(\mathbf{x}_i)} \quad (\text{S28})$$

| | Recurrent step | Gradient step |
|----------|---|--|
| (E-step) | Hold $\pi_{i,k}^{(t-1)}$ and $\boldsymbol{\theta}_k^{(t-1)}$ constant for all i | $\frac{\partial \mathcal{O}_i}{\partial \mathcal{Q}}$; find local maximum at $q^{\dagger(t)}$ |
| (M-step) | Hold $q^{\dagger(t)}$ constant | $\frac{\partial \mathcal{O}_i}{\partial \boldsymbol{\theta}}$; find local maximum |

Table 1: EM framed as gradient ascent^{21,23,28}. This framing shows how different pixels each contribute to increasing the overall data likelihood as long as there is information in q^\dagger about neighbors. The update rule in equation 2 is derived by finding closed-form solutions for local maxima when setting the gradient steps equal to 0 and then applying the function u . Values for $\pi_{i,k}^{(t)}$ are updated until the total likelihood $\log p(\mathbf{X} | \mathbf{\Pi}, \boldsymbol{\theta})$ reaches an asymptote.

We write the denominator as the marginalized sum over all classes. We also introduce the symbol $\gamma_{i,k}^{(t)}$, termed the “responsibility” which quantifies how “responsible” a particular class is for a particular observation.

$$\gamma_{i,k}^{(t)} \equiv \frac{p(\mathbf{x}_i | c_{i,k}, \boldsymbol{\theta}_k^{(t-1)}) p(c_{i,k} | \pi_{i,k}^{(t-1)})}{\sum_k^K p(\mathbf{x}_i | c_{i,k}, \boldsymbol{\theta}_k^{(t-1)}) p(c_{i,k} | \pi_{i,k}^{(t-1)})} \quad (\text{S29})$$

Now we can substitute equation S29 into equation S27:

$$\mathcal{O}_i(q^\dagger^{(t)}, \boldsymbol{\vartheta}) = \sum_k^K (\log p(\mathbf{x}_i, c_{i,k} | \pi_{i,k}, \boldsymbol{\theta}_k)) \cdot \frac{p(\mathbf{x}_i | c_{i,k}, \boldsymbol{\theta}_k^{(t-1)}) p(c_{i,k} | \pi_{i,k}^{(t-1)})}{\sum_k^K p(\mathbf{x}_i | c_{i,k}, \boldsymbol{\theta}_k^{(t-1)}) p(c_{i,k} | \pi_{i,k}^{(t-1)})} \quad (\text{S30})$$

$$= \sum_k^K (\log p(\mathbf{x}_i, c_{i,k} | \pi_{i,k}, \boldsymbol{\theta}_k)) \cdot \gamma_{i,k}^{(t)} \quad (\text{S31})$$

Decompose the joint probability distribution

$$= \sum_k^K (\log(p(\mathbf{x}_i | c_{i,k}, \boldsymbol{\theta}_k) \cdot p(c_{i,k} | \pi_{i,k}))) \cdot \gamma_{i,k}^{(t)} \quad (\text{S32})$$

$$= \sum_k^K [\log(\text{Student}(\boldsymbol{\theta}_k)) + \log(\pi_{i,k})] \cdot \gamma_{i,k}^{(t)} \quad (\text{S33})$$

Here we show the maximization with respect to $\pi_{i,k}$ (see Peel¹² for estimation of the parameters $\boldsymbol{\theta}_k$). Maximization for $\pi_{i,k}$ is subject to the constraint $\sum_k^K \pi_{i,k} = 1$, implemented with a Lagrange multiplier λ :

$$\mathcal{O}_i(q^\dagger^{(t)}, \pi_{i,k}, \lambda) = \gamma_{i,k}^{(t)} \cdot \log(\pi_{i,k}) + \lambda \left(1 - \sum_k^K \pi_{i,k}\right) \quad (\text{S34})$$

$$\frac{\partial \mathcal{O}_i}{\partial \pi_{i,k}} = \frac{\gamma_{i,k}^{(t)}}{\pi_{i,k}} - \lambda \quad (\text{S35})$$

Set the partial derivative equal to 0 and solve for $\pi_{i,k}$ to find the updated value $\pi_{i,k}^{(t)}$:

$$\lambda = \frac{\gamma_{i,k}^{(t)}}{\pi_{i,k}} \quad (\text{S36})$$

$$\pi_{i,k}^{(t)} = \frac{\gamma_{i,k}^{(t)}}{\lambda} \quad (\text{S37})$$

Equation S37 is the standard formulation of the EM update rule, because $\sum_k^K \pi_{i,k}^{(t)} = 1$ we can write that $\lambda = \sum_k^K \gamma_{i,k}^{(t)}$. Our formulation differs however because of the u function applied to the complete log-likelihood so we write:

$$\lambda = \frac{u_i(\gamma_{\cdot,k}^{(t)})}{\pi_{i,k}} \quad (\text{S38})$$

Finally, this leads to our update rule:

$$\pi_{i,k}^{(t)} = \frac{u_i(\gamma_{\cdot,k}^{(t)})}{\sum_k^K u_i(\gamma_{\cdot,k}^{(t)})} \quad (\text{S39})$$

Xu and Jordan²⁸ have shown that EM can be considered a quasi-Newton’s method, which is a classical iterative optimization method.

We end with a summary of how our approximate distribution q changes with time. We started with an approximate distribution over \mathbf{c} , $\mathbf{c} \sim q$. We show that to maximize likelihood/minimize free-energy $q \rightarrow q^\dagger \equiv p(\mathbf{c} | \mathbf{x}) \equiv \gamma^{(t)}$ which is then rescaled to $\pi^{(t)}$ and that through spatial smoothing with u , $\boldsymbol{\pi}$ has further, single-pixel dynamics π_i .

D.5. Deep spatial prior. As explained in Section A, we use the notation $\mathbf{x}^{[l]}$ to indicate that our random variable formulation of the observations \mathbf{X} is in a subspace of $\mathbf{Z}^{[l]}$ for a specific layer $[l]$. This leads to latent variables being layer specific as well (e.g. $\mathbf{c}^{[l]}$, which was omitted in earlier sections for clarity). Just as class labels \mathbf{c} that are close together in space are more likely to be the same, we assumed that so too should class labels for the same location across neighboring layers $\{\mathbf{c}^{[l-1]}, \mathbf{c}^{[l]}, \mathbf{c}^{[l+1]}\}$. This leads to better segmentation performance with FlexMM⁶.

Therefore, we can write equation S24 for a spatial neighborhood while considering neighboring layers as well:

$$\pi_{i,k}^{(t)[l]} = \frac{u_i \left(\gamma_{\cdot,k}^{(t)[l-1]}, \dots, \gamma_{\cdot,k}^{(t)[l+1]} \right)}{\sum_k^K u_i \left(\gamma_{\cdot,k}^{(t)[l-1]}, \dots, \gamma_{\cdot,k}^{(t)[l+1]} \right)} \quad (\text{S40})$$

The appropriate linear function u for applying this deep spatial prior and the corresponding update rule is presented in Box I of our FlexMM paper⁶.

D.6. EM initial guess. For our initial guesses, $\pi_{i,k}^{(t=0)}$ and $\theta_k^{(t=0)}$, first we define $\pi_{i,k}^{(0)}$ by sampling from resized, probabilistic segmentation maps of the participants (Fig. 1b, top). (Resizing is necessary because the coordinate system for participant segmentation maps is lower resolution than the per-pixel coordinate system of FlexMM). Let the value at each pixel in this resized participant segmentation map be notated \mathbf{p} (analogous to $\boldsymbol{\pi}$ from FlexMM, that we have discussed in detail, but based on participant responses). We can then define sampled labels:

$$s_i \sim \text{Mult}(\mathbf{p}) \quad (\text{S41})$$

We then compute $\pi_{i,k}^{(t=0)}$ from s_i like so.

$$\pi_{i,k}^{(t=0)} = \begin{cases} 1 - \epsilon & \text{if } k = s_i \\ 0 + \epsilon / (K - 1) & \text{if } k \neq s_i \end{cases} \quad (\text{S42})$$

After computing this mapping, we estimate $\mu_k^{(0)}$ and $\boldsymbol{\Sigma}_k$ using assignments s_i . For the initial guess of ν we use $\nu = 2$, which simplifies the mixer distribution $\mathbb{P}(\nu)$ to a scaled inverse Rayleigh distribution^{6,11}.

Note that by sampling in the initial guess, our model is sensitive to different levels of perceptual uncertainty in each case (*i.e.* each participant and image). This is part of the reason we refer to the EM initial guess as the participant's mental map in the main text. While the final maps from FlexMM are different depending on the precise sample of initial guess, our main findings on dynamics are qualitatively robust across varied initial guesses.

E. Graphical model perturbations. In this section we focus on the ablations detailed in Figure 5.

E.1. No local connectivity model. If we define the u function as a Gaussian with an infinitely wide kernel *i.e.* $u(\gamma_{N,k}^{(t)}) = \sum_{i=1}^N \gamma_{i,k}^{(t)}$, where $N = h \times w$, then we can plug back into equation S29 to get:

$$\pi_{i,k}^{(t)} = \frac{\sum_{i=1}^N \gamma_{i,k}^{(t)}}{\sum_{i=1}^N \sum_{k=1}^K \gamma_{i,k}^{(t)}} \quad (\text{S43})$$

By definition, $\sum_{i=1}^N \sum_{k=1}^K \gamma_{i,k}^{(t)} = N$ so:

$$\pi_k^{(t)} = \frac{\sum_{i=1}^N \gamma_{i,k}^{(t)}}{N} \quad (\text{S44})$$

Because we are summing over all i , the index i is no longer present on the left-hand-side of equation S44. Therefore, to obtain per-pixel probabilities, we have to compute the posterior (which is also equivalent to the E step of the next iteration):

$$q_{i,k}^{(t)} = p \left(\mathbf{c} = k \mid \pi_k^{(t)}, \theta_k^{(t)} \right) \quad (\text{S45})$$

Further Discussion:

Because the no local connectivity model yielded similar performance to the full IBI model (Fig. 5), we take a closer look at the nature of the no local connectivity perturbation.

There are two redundancies in our algorithm which contribute to stabilizing performance in the no local connectivity model. First, one such redundancy is using the human subjective map, which contains spatially smooth segments, as the initial guess. This ensures that $\gamma_{i,k}^{(0)}$ in equation S43 is already smoothed, which influences the $q_{i,k}^{(t)}$ at later iterations in equation S45. Indeed, when we considered a no local connectivity model with random initialization, it did not converge to meaningful segmentation maps. This further corroborates that the initialization to a smooth map contributes to obtaining a smooth map at convergence.

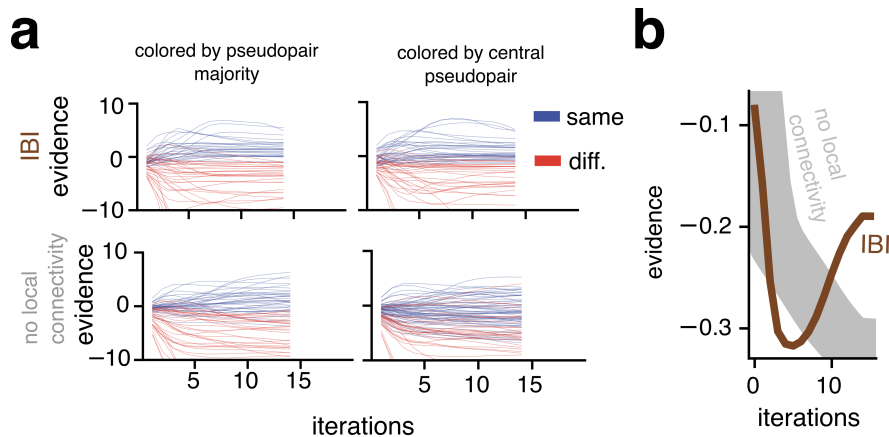
The second redundancy is the fact that the model does not output the reaction time or choice for a single pair of pixels. We average the model reaction time and response at pair i, j over multiple pseudopairs, namely pairs of coordinates that are close to i and j (see Methods). Averaging over pseudopairs introduces some spatial smoothing, reducing the difference between the full IBI model and the no local connectivity model. Nonetheless, we can use these pseudopairs to compare the two models in greater depth. We introduce a new quantity we call spatial confidence, which quantifies the agreement across pseudopairs for a certain choice. For example, if 51% of pseudopairs around pixels x and y yield positive evidence this would lead the model to output a “same segment” response for pixels x, y . If, on the other hand, 99% of pseudopairs around pixels w and z have positive evidence there is clearly more spatial confidence for the w, z pair than the x, y pair, but the model outputs the same binary “same segment” response. As expected from the lack of spatial prior in the no local connectivity model, we found that its spatial confidence was lower than the IBI model (Table 2).

| | IBI | no local connectivity |
|--------------------|------|-----------------------|
| same segment | 0.87 | 0.79 |
| different segments | 0.95 | 0.89 |

Table 2: Comparing spatial confidence for the IBI model and the no local connectivity model. Left-column values are higher than right-column values in both rows. The extent to which these number are still quite similar is due to the initial guess, so it stands to reason that the local-connectivity prior improves on spatial confidence.

This choice to have the majority of pseudopairs yield a response stabilizes the no local connectivity model in a way that would be perturbed if we were to, say, consider only the response of the central pseudopair. The full IBI model is not susceptible to this perturbation (Supp. Fig. 14a).

To further elucidate the difference between IBI and the no local connectivity model, we performed a temporal PCA where the evidence trace for each pseudopair was treated as an observation and each time-step was treated as a feature to find temporal basis functions for each model. We found that the no local connectivity model’s traces displayed one order of magnitude more variance around the first-component basis function compared to the full IBI model. We also found that IBI’s first-component basis function was more complex and showed oscillatory behavior (Supp. Fig. 14b).



Supplementary Figure 14: The no local connectivity model is not robust to small spatial perturbations. **a**, Top row: evidence traces for the full IBI model colored by the pseudopair majority (left), versus the central pseudopair (right). The two partitions are almost equivalent. Bottom row: evidence traces for the no local connectivity model colored by the pseudopair majority (left) versus the central pseudopair (right). The partition coloring by the central pseudopair shows many same segment traces yielding negative evidence, because evidence from surround pixels is not in accordance with center pixels. **b**, EM iterations (time) are plotted on the x -axis, while evidence is plotted on the y -axis. Note that the amplitude of the evidence is arbitrary here as the actual evidence is a scalar multiple of these plotted basis functions. The thickness of each plotted line indicates the variance of evidence traces around the basis function. In the no local connectivity model, there is 10 times more variance than in the IBI model.

Ultimately these differences may be obscured by averaging over pseudopairs to predict behavioral proxies such as reaction time, but more work could be done to develop testable hypothesis about spatial confidence and variability and determine which model is better aligned to the empirical data.

E.2. No mental map model. For the random mental map we sampled the initial guess from a multinomial distribution where samples had equal probability of segment label up to the prescribed number of segments K :

$$c_{i,k}^{(t=0)} \sim \text{Mult}[(1/K, \dots, 1/K)_K] \quad (\text{S46})$$

$$\pi_{i,k}^{(t=0)} = \begin{cases} 1 - \epsilon & \text{if } k = c_{i,k} \\ 0 + \epsilon/(K-1) & \text{if } k \neq c_{i,k} \end{cases} \quad (\text{S47})$$

F. Model convergence. Having defined the initial state of our algorithm (the “mental map”), and the processing at each time step we now define model convergence.

F.1. The iterative timescale. As parameters are updated at each time step, $\boldsymbol{\pi}^{(t)}$ and $\boldsymbol{\theta}^{(t)}$ can be used to compute the value of the observation log-likelihood per time-step $\log p(\mathbf{X} | \boldsymbol{\Pi}^{(t)}, \boldsymbol{\theta}^{(t)})$ shortened $\mathcal{L}^{(t)}$. We define the time of convergence, t^*

$$t^* : \mathcal{L}^{(t)} - \mathcal{L}^{(t-1)} < \delta \mathcal{L} \quad (\text{S48})$$

Where $\delta \mathcal{L}$ is a small, positive constant.

Therefore, we notate the per-pixel segment probabilities at convergence as $\pi_{i,k}^{(t^*)}$. Conceptually, we can think of t^* as the time at which further iteration does not meaningfully increase observation likelihood. In other words, after time t^* there is no informational benefit to the computational cost of iteration.

G. Evidence for “same segment”. Thus far, we have described FlexMM which provides us with $\pi_{i,k}^{(t)}$, *i.e.* the segment probabilities for each pixel, up until convergence at $\pi_{i,k}^{(t^*)}$ to get from these to evidence about the “same segment” hypothesis we rewrite equations 3–4 from the main text, Methods.

$$\pi_{ij}^{(t)} = \sum_k^K \pi_{i,k}^{(t)} \pi_{j,k}^{(t)} \quad (\text{S49})$$

$$\mathbf{E}_{ij}(t) = \left[\log \left(\pi_{ij}^{(t)} \right) - \log \left(1 - \pi_{ij}^{(t)} \right) \right] \quad (\text{S50})$$

π_{ij} in equation S49 is the probability that two pixels i and j belong to the same segment. From this probability we use the log-likelihood ratio as evidence, notated \mathbf{E}_{ij} . Note that although \mathbf{E} is not a vector quantity, we use boldface to emphasize that \mathbf{E} is generated by the IBI model using multi-dimensional inference procedures. The use of the log-likelihood ratio is a principled choice in models of decision-making, including the DDM, as will be further discussed in the upcoming Section 12.

Supplementary Note 12: Drift Diffusion modeling

Previous work in decision modeling states that an optimal observer—that is one who responds the fastest given an acceptable maximum error rate—can be modeled using the drift diffusion model (DDM) ²⁹. The goal of a DDM is to predict reaction time distributions per choice–type ³⁰. In this work, we do fit DDMs to reaction time distributions, but our goal is not prediction. Instead we wish to compare the widely-adopted stochastic linear dynamics of a DDM to the nonlinear dynamics of IBI. For fair comparison, we developed an image-computable DDM and here we review the assumptions we followed in order to do so.

A. Sequential log-probability ratio test. First, we define the fundamental assumptions of DDMs. DDMs model decision-making as a sequential analysis problem, in which evidence, ξ , about the stimulus is accumulated over s sequential samples (*i.e.* $\xi_0, \xi_1, \dots, \xi_s$). The optimal way to consider this evidence is using the sequential probability ratio test ^{29,31–33}, where h_y is the hypothesis that two cues are in the same segment while h_n is the hypothesis that two cues are in a different segment:

$$L_s = \log \left(\frac{p(\xi_{0 \rightarrow s} | h_y)}{p(\xi_{0 \rightarrow s} | h_n)} \right) \quad (\text{S51})$$

Assuming that samples are independent we can take the limit approaching a differential amount of time (dt):

$$\lim_{s \rightarrow dt} (L_s) \equiv \ell = \log \left(\frac{p(\xi_{dt} | h_y)}{p(\xi_{dt} | h_n)} \right) \quad (\text{S52})$$

In DDMs, a simplifying assumption is made that because evidence is drawn from a large population of neurons perturbed by noise, the central limit theorem applies, and the differential log–likelihood can be re-parametrized as a Gaussian with mean m and variance η^2 (see Bogacz ²⁹ Appendix A for a more detailed derivation).

$$\ell = m + \eta dW_t \quad (\text{S53})$$

Classical DDMs use a stochastic process called the Wiener process to model noise. The Wiener process $W(t)$ is introduced such that $dW \sim \sqrt{dt} \cdot \mathcal{N}$ where \mathcal{N} is a Gaussian with zero mean and unit variance. Following from the independent samples assumption made earlier: for $0 < s < t < u < v$, $W(v) - W(u)$ is independent from $W(t) - W(s)$ ²⁹. Using the Wiener process we can reparametrize log-likelihood ratio L as a function of time $L(t)$:

$$L(t) = z + m \int_0^t ds + \eta \int_0^t dW_s \quad (\text{S54})$$

Where z is introduced as a starting point bias. Integral notation makes clear that evidence accumulation (a.k.a. evidence integration) refers to averaging out noise over time, but here on we write $L(t)$ without integrals for simplicity, like so:

$$L(t) = z + mt + \eta W(t) \quad (\text{S55})$$

B. Iterative and Integrative timescales. Treatment of the differential log–likelihood ℓ is where FlexMM differs from classical DDMs. In FlexMM, the dynamics of ℓ is determined by the information that FlexMM extracts from the image, and thus ℓ is affected by the connectivity of the graph. Comparing $L(t)$ to $\mathbf{E}_{ij}(t)$ from equation S50 in Supplementary Note 11, we can appreciate that they are both log-probability ratios. Thanks to this shared probabilistic formulation, we were able to further develop bespoke DDMs that leverage image statistics while eliminating the nonlinear within-trial dynamics of iteration in IBI.

The intuitive approach was to assume that we can substitute the parameters in $L(t)$ with value of the *unchanging* log-likelihood at iteration convergence $\mathbf{E}_{ij}(t^*)$, divided by t^* . We introduce a new variable for differential evidence in the integrative timescale:

$$\mathcal{E}_{ij} = \frac{\mathbf{E}_{ij}^{(t^*)}}{t^*} \quad (\text{S56})$$

This quotient is the average value of sensory evidence during the iterative timescale. Notionally, we divide by t^* to keep our units consistent with m . Practically, this division ensures that stimuli that take longer to converge do not have their evidence count for more, which would be counter–intuitive.

Therefore, technically speaking, $L(t)$ is only defined for $t > t^*$. Following the assumptions of sequential analysis, this creates a decision-making model in which the integrative timescale follows the iterative timescale. The model which yields a better fit is deemed to “set the tempo”. In other words, it is a better model of the slow step of the decision-making process. That is, if DDM is a better model, extracting the evidence from the input image either takes the same amount of time for all pairs or is much faster (and thus has less influence on reaction times) than the process of averaging out noise. Conversely, if IBI is a better model, reaction times are largely determined by the dynamic process that extracts evidence from the sensory input, whereas additional noise is negligible. See Table 3 for a side-by-side comparison of timescales.

Note that neither $\mathbf{E}_{ij}(t)$ nor $L_{ij}(t)$ are closed-form continuous expressions. However, both the models can be thought of as using a first-order numerical integration procedure to approximate a continuous evidence function.

IBI (for $t \rightarrow t^*$):

$$\mathbf{E}_{ij}(t) = \left[\log \left(\pi_{ij}^{(t)} \right) - \log \left(1 - \pi_{ij}^{(t)} \right) \right]$$

Decision rules:

$$\begin{aligned} \rho_1 &= \{0 \leq t \leq t^* : |\mathbf{E}_{ij}(t)| \geq b\} \\ \rho_2 &= \left\{ 0 \leq t \leq t^* : \left| \frac{\partial}{\partial t} (\mathbf{E}_{ij}) \right| \leq a |\mathbf{E}_{ij}| \right\} \end{aligned} \quad (\text{S57})$$

Fittable Parameters (ψ):

$$\psi = (a, b)$$

Decision time distributions:

$$\hat{\mathbf{t}}_{d,y} = \inf\{(\rho_1 \cup \rho_2) \cap (\mathbf{E}_{ij}(t) > 0)\} \quad (\text{S58})$$

$$\hat{\mathbf{t}}_{d,n} = \inf\{(\rho_1 \cup \rho_2) \cap (\mathbf{E}_{ij}(t) < 0)\} \quad (\text{S59})$$

DDMs (for $t^* \rightarrow T$):

$$L_{ij}(t) = z + \alpha \cdot m_{ij}t + \eta W(t)$$

Decision rules:

$$\rho_3 = \{t^* \leq t \leq T : |L_{ij}(t)| \geq \beta\}$$

Fittable Parameters (ψ):

$$\psi = (\alpha, \beta)$$

Decision time distributions:

$$\hat{\mathbf{t}}_{d,y} = \inf\{(\rho_3) \cap (L_{ij}(t) > 0)\} \quad (\text{S60})$$

$$\hat{\mathbf{t}}_{d,n} = \inf\{(\rho_3) \cap (L_{ij}(t) < 0)\} \quad (\text{S61})$$

Table 3: A side-by-side comparison of the iterative and integrative timescales

C. Types of DDMs. All the DDMs we use can be defined with the following parameterization, where $W_{ij}(t)$ indicates that each unique pair (i, j) is an independent draw from the Wiener process:

$$L_{ij} = z_{ij} + \alpha \cdot m_{ij} + \eta W_{ij}(t) \quad (\text{S62})$$

See Table 4 for how we modify substitutions of \mathcal{E}_{ij} to create the different DDMs reported in our main text as well as in Supplementary Note 6.

| z_{ij} | m_{ij} | Heuristic | Nomenclature in main text |
|--------------------|--|---------------------------------|---|
| 0 | 1 | α is decoded from data | Base DDM |
| 0 | $\langle \mathcal{E}_{ij} \rangle_{\hat{y}^*} / \hat{n}^*$ | weight by choice-type | choice-weighted DDM |
| \mathcal{E}_{ij} | \mathcal{E}_{ij} | weight per trial | image-computable DDM/trial-weighted DDM |
| 0 | \mathcal{E}_{ij} | weight drift rate per trial | reduced trial-weighted DDM |
| \mathcal{E}_{ij} | $\langle \mathcal{E}_{ij} \rangle_{\hat{y}^*} / \hat{n}^*$ | weight starting point per trial | reduced trial-weighted DDM |

Table 4: Types of DDMs. The angle-bracket notation indicates an average. The notation \hat{y}^* / \hat{n}^* indicates that evidence is averaged per the choice type of the model, at convergence; y : same segment, n : different segments.

Supplementary Note 13: Parameter fitting

In order for the model to generate predictions $\hat{\mathbf{t}}_d$ that are most aligned with empirical observations \mathbf{t}_d , we implemented a parameter fitting procedure. Only two parameters, $\psi = a, b$ for IBI or $\psi = \alpha, \beta$ for DDM, were used for the fitting while the noise η remained fixed.

To be able to compare the model's output of reaction time (which is based on an iteration index) with the continuous reaction time (based on physical time) from humans performing the task, we apply the following transformation to both model and human reaction times: $\mathbf{t}'_d = \text{resc}(\log(\mathbf{t}_d + 1))$. Where the $\text{resc}(\mathbf{x})$ function is simply $(\mathbf{x} - \mathbf{x}_{\min})/(\mathbf{x}_{\max} - \mathbf{x}_{\min})$.

The goal is to maximize the likelihood that a human decision time \mathbf{t}'_d comes from the distribution of model reaction times $\hat{\mathbf{t}}'_d$. State-of-the-art DDMs use an analytical solution to construct the PDF and compute a loss $\hat{f}(\hat{\mathbf{t}}'_d|\psi)$ ^{34,35}. While this would be possible for the base DDM model, it is not applicable to all models so we opted to empirically construct CDFs for loss instead: $\hat{F}(\hat{\mathbf{t}}'_d|\psi)$. We can use the following empirical likelihood function to calculate the likelihood that a distribution of human reaction times \mathbf{t}'_d comes from the model CDF \hat{F} ³⁶:

$$p(\mathbf{t}'_d|\hat{F}, \psi) = \prod_{n=1}^{N_T} \frac{\hat{F}(t'_{d,n}|\psi) - \hat{F}(t'_{d,n} - \delta(t'_{d,n})|\psi)}{\delta(t'_{d,n})} \quad (\text{S63})$$

where the product is over all trials, and $\delta(t'_{d,n})$ is the time difference with the closest observation smaller than $t'_{d,n}$. For numerical stability, we optimize the negative log-likelihood $-\mathcal{L}(\mathbf{t}'_d, \hat{F}, \psi)$.

$$-\mathcal{L}(\mathbf{t}'_d, \hat{F}, \psi) = \sum_{n=1}^{N_T} \log \left[\hat{F}(t'_{d,n}|\psi) - \hat{F}(t'_{d,n} - \delta(t'_{d,n})|\psi) \right] - \log(\delta(t'_{d,n})) \quad (\text{S64})$$

The second term is not a function of the parameters so we ignore it. For the final loss function, we apply the above equation independently to each set of choices

$$-\mathcal{L}_{\text{total}} = -\mathcal{L}(\mathbf{t}'_{d,y}, \hat{F}_y, \psi) - \mathcal{L}(\mathbf{t}'_{d,n}, \hat{F}_n, \psi) \quad (\text{S65})$$

Note that the subscripts y and n indicate the “same segment” subset, and the “different segments” subset as defined by human choices, ignoring the choices of the model. This is because the choices made by the model at specific trials were not expected to match with the human's, the objective was simply to match reaction time distributions per choice type. Equation S65 is not differentiable with respect to the parameters. Therefore we used Markov-Chain Monte-Carlo (MCMC) global minimization methods^{37,38}. Starting from an initial guess (e.g., for IBI, $\psi^{(0)} = a^{(0)}, b^{(0)}$) we compute the negative log-likelihood $-\mathcal{L}_{\text{total}}$, which is then iteratively minimized. We heuristically set the following bounds for the MCMC optimizer on parameters in ψ : $a \in (0, 1]$, $b \in (0, 5]$, $\alpha \in (0, 50]$, $\beta \in (0, 5]$. We also found that the minimization of loss improved when multiple initial guesses of ψ were used. We used two different initial guesses.

By computing this minimization we are finding parameters that maximize the likelihood of human decision times being predicted by our model. For fair comparison across models we cross-validated each model fit with a 5-fold cross validation per case.

Supplementary Note 14: IBI model extensions

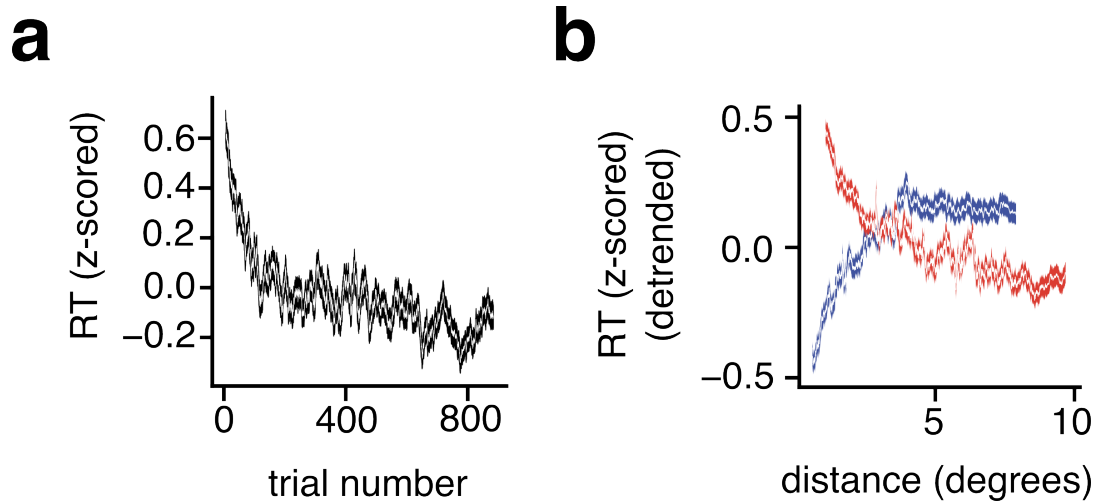
The choices made in our modeling (e.g. using a convolutional neural network for feature extraction or learning parametric distributions) are rooted in existing literature, and allow for controlled model comparison with alternatives that shed light on exactly what aspects of the model are useful for capturing the spatiotemporal dynamics of human segmentation. Our model lays the foundation however, for a number of extensions which could potentially improve the quantitative performance of IBI to predict reaction times, detailed below.

Extending feature extraction. We use the deep convolutional neural network VGG-19 for feature extraction because it has a receptive field like architecture that mimics early visual cortex^{3,4}. In computer vision however, the field has moved away from deep convolutional neural networks and towards more powerful transformer models from which segmentation features which include semantic information may emerge³⁹. These segmentation features can easily be plugged into our iterative generative model which could then predict reaction times using iteration index.

Extending decision rules and parameters to fit reaction time data. We used the simplest possible decision rules given our goal of comparing to DDMS, which led our IBI model to have only two fittable parameters to fit reaction times. Given there are hundreds of trials, we could add decision rules and new parameters. One example of a new decision rule would be to take into account the trial number, *i.e.* the sequential order in which pairs of cues are presented during the experiment (see Supp. Fig. 15).

Finally, consistent with the DDM literature, our objective function is minimized to match the overall reaction time distribution rather than predicting reaction times per trial. Instead of matching empirical CDFs, we could use a more standard loss function, for example the squared error.

Supplementary Note 15: Reaction times decrease throughout the experimental block



Supplementary Figure 15: **a**, z-scored reaction time vs. ordered trial number. Trials are aggregated across all cases. The white line represents the mean within a sliding window of 600 trials. Black shaded area: s.e.m. **b**, Same as Figure 2a, bottom, with detrended data (*i.e.* removing the mean trend of panel **a**; details in the text below).

We observe that RTs decrease as the task goes on (Supp. Fig. 15a). To ensure that the presence of this trend (which is unrelated to the distance between cues) did not affect the RT correlations with distance, we detrended the RT data and verified that results did not change. To detrend the data, we fit a decaying exponential to the z-scored RT vs. trial number, and then subtracted the predictions made by this exponential from the empirical z-scored RT values. Supp. Fig. 15b shows that this detrending procedure did not change our findings.

References

1. Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* **2**, 193–218 (1985).
2. Ruscio, J. Constructing confidence intervals for spearman’s rank correlation with ordinal data: A simulation study comparing analytic and bootstrap methods. *J. Mod. Appl. Stat. Methods* **7**, 416–434 (2008).
3. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations* (2015).
4. Cadena, S. A. *et al.* Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Comput. Biol.* **15**, e1006897 (2019).
5. Schrimpf, M. *et al.* Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron* **108**, 413–423 (2020).
6. Vacher, J., Launay, C. & Coen-Cagli, R. Flexibly regularized mixture models and application to image segmentation. *Neural Netw.* **149**, 107–123 (2022).
7. Wainwright, M. J. & Simoncelli, E. Solla, S., Leen, T. & Müller, K. (eds) *Scale Mixtures of Gaussians and the Statistics of Natural Images.* (eds Solla, S., Leen, T. & Müller, K.) *Advances in Neural Information Processing Systems*, Vol. 12 (MIT Press, 1999).
8. Schwartz, O. & Simoncelli, E. P. Natural signal statistics and sensory gain control. *Nat. Neurosci.* **4**, 819–825 (2001).
9. Coen-Cagli, R., Dayan, P. & Schwartz, O. Cortical surround interactions and perceptual salience via natural scene statistics. *PLoS Comput. Biol.* **8**, e1002405 (2012).
10. Coen-Cagli, R. & Mamassian, P. Are we ready to tackle perceptual segmentation of natural scenes? *Vision Res.* **240**, 108749 (2025).
11. Festa, D., Aschner, A., Davila, A., Kohn, A. & Coen-Cagli, R. Neuronal variability reflects probabilistic inference tuned to natural image statistics. *Nat. Commun.* **12**, 3635 (2021).
12. Peel, D. & McLachlan, G. J. Robust mixture modelling using the t distribution. *Stat. Comput.* **10**, 339–348 (2000).
13. Koller, D. & Friedman, N. *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, 2009).
14. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the *em* algorithm. *J. R. Stat. Soc. Series B Stat. Methodol.* **39**, 1–22 (1977).
15. Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)* 1 edn (Springer, 2007).
16. Eckford, A. W. The factor graph EM algorithm: applications for LDPC codes. *IEEE 6th Workshop on Signal Processing Advances in Wireless Communications, 2005* 910–914 (2005).
17. Dauwels, J., Eckford, A. W., Korl, S. & Loeliger, H. Expectation maximization as message passing - part I: principles and gaussian messages. *CoRR* **abs/0910.2832** (2009). URL <http://arxiv.org/abs/0910.2832>.
18. Lee, T. S. & Mumford, D. Hierarchical bayesian inference in the visual cortex. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **20**, 1434–1448 (2003).
19. Beck, J. M., Latham, P. E. & Pouget, A. Marginalization in neural circuits with divisive normalization. *J. Neurosci.* **31**, 15310–15319 (2011).
20. Raju, R. V., Li, Z., Linderman, S. & Pitkow, X. Inferring inference. *CoRR* **abs/2310.03186** (2023). URL <https://doi.org/10.48550/arXiv.2310.03186>.
21. Neal, R. M. & Hinton, G. E. in *A view of the em algorithm that justifies incremental, sparse, and other variants* (ed. Jordan, M. I.) *Learning in Graphical Models*, Vol. 89 of *NATO ASI Series* 355–368 (Springer Netherlands, 1998).
22. Roweis, S. & Ghahramani, Z. A unifying review of linear gaussian models. *Neural Comput.* **11**, 305–345 (1999).
23. Salakhutdinov, R., Roweis, S. T. & Ghahramani, Z. Optimization with EM and expectation-conjugate-gradient. *ICML’03: Proceedings of the Twentieth International Conference on International Conference on Machine Learning* 672–679 (2003).
24. Friston, K. A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **360**, 815–836 (2005).
25. Buxó, C. E. R. & Savin, C. Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P. & Vaughan, J. W. (eds) *A sampling-based circuit for optimal decision making.* (eds Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P. & Vaughan, J. W.) *Advances in Neural Information Processing Systems*, 14163–14175 (2021).
26. Chen, S., Jiang, L., Rao, R. P. N. & Shea-Brown, E. Oh, A. *et al.* (eds) *Expressive probabilistic sampling in recurrent neural networks.* (eds Oh, A. *et al.*) *Advances in Neural Information Processing Systems* (2023).
27. Zhang, W.-H., Wu, S., Josić, K. & Doiron, B. Sampling-based bayesian inference in recurrent circuits of stochastic spiking neurons. *Nat. Commun.* **14**, 7074 (2023).

28. Jordan, M. I. & Xu, L. Convergence results for the EM approach to mixtures of experts architectures. *Neural Netw.* **8**, 1409–1431 (1995).
29. Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J. D. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* **113**, 700–765 (2006).
30. Fudenberg, D., Newey, W., Strack, P. & Strzalecki, T. Testing the drift-diffusion model. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 33141–33148 (2020).
31. Neyman, J. & Pearson, E. S. IX. on the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Lond.* **231**, 289–337 (1933).
32. Wald, A. *Sequential analysis* (John Wiley, 1947).
33. Wald, A. & Wolfowitz, J. Optimum character of the sequential probability ratio test. *Ann. Math. Stat.* **19**, 326–339 (1948).
34. Feller, W. *An Introduction to Probability Theory and Its Applications* Vol. 1 (Wiley, 1968).
35. Navarro, D. J. & Fuss, I. G. Fast and accurate calculations for first-passage times in wiener diffusion models. *J. Math. Psychol.* **53**, 222–230 (2009).
36. Owen, A. *Empirical Likelihood* (CRC Press, 2001).
37. Wales, D. J. & Doye, J. P. K. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J. Phys. Chem. A* **101**, 5111–5116 (1997).
38. Xiang, Y., Sun, D. Y., Fan, W. & Gong, X. G. Generalized simulated annealing algorithm and its application to the thomson model. *Phys. Lett. A* **233**, 216–220 (1997).
39. Oquab, M. *et al.* Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.* **2024** (2024).