A single inference process underlies the behavior and neural circuits of working memory

Dongyu Gong (dongyu.gong@yale.edu)[1], Mario Belledonne (mario.belledonne@yale.edu)[1], and Ilker Yildirim (ilker.yildirim@yale.edu)[1,2]

[1]Department of Psychology, Yale University
[2]Wu Tsai Institute, Yale University

Abstract

Working memory holds past inputs online for future behavior. Descriptive models characterize what limits its capacity, and mechanistic models specify how neurons maintain it. However, despite decades of work, we are missing an algorithmic account that both explains the behavioral signatures of working memory and can be realized in a neural circuit. Here we propose MemGP, an algorithmic framework that derives both from a single inference process. MemGP represents working memory as a Gaussian-process posterior over a memory-strength function on feature space, and encoding, maintenance and retrieval are operations on this posterior. In simulated experiments, MemGP reproduces load-dependent precision errors, attentional benefits during maintenance, and attraction and repulsion between remembered items. The objective that MemGP optimizes also specifies a recurrent circuit in which the posterior mean corresponds to population activity, the covariance to a rapidly updated state, and the prior to the center–surround connectivity assumed by attractor models. This single inference objective thus accounts for working-memory behavior and specifies candidate circuit dynamics, providing the algorithmic account that has been missing between descriptive and mechanistic models. MemGP therefore makes both behavioral and neural predictions and offers a general framework that could extend to perception, long-term memory and decision making.

1 Introduction

Working memory is a cornerstone of human cognition, making past inputs available for future behavior by holding information temporarily online. Models of working memory have been developed at both descriptive and mechanistic levels (Fig. 1b). Descriptive models characterize what limits it, whether a fixed number of slots [1, 2], or a continuous resource [3–5]. Mechanistic models specify how neurons maintain it, through persistent activity [6–8] or short-term changes in synaptic efficacy [9, 10]. Both traditions have enjoyed some success [11, 12], and several attempts have been made to reconcile competing accounts [13–18]. These efforts, however, address mainly the capacity limit of working memory. A more basic problem is that the two traditions share no common

representation: we are missing an algorithmic framework that derives the content of working memory (what to store) and its neural implementation (where to store) from the same principle.

Such a framework should provide a representation that both descriptive and mechanistic models can share. It should define the resource that supports the representation and the operations that can act on the representation. This representation should also jointly store the content and uncertainty of working memory [19–21], so that downstream decisions can take the reliability of the memory into account. The operations on the representation should map onto a biologically plausible circuit, so that resource allocation among items can be implemented by neural dynamics or synaptic changes. This framework would have broad consequences. It would unify the currently separate models of working memory, and its predictions could be tested against both behavioral and neural data. It could also extend to perception, long-term memory and decision making, relating working memory to other aspects of cognition within one theory.

Here we propose MemGP, an algorithmic framework that meets these requirements. MemGP represents working memory as a posterior distribution over a memory-strength function on feature space. The function assigns a strength of evidence to each feature combination, such as a particular color at a particular location, and a Gaussian process (GP) describes the distribution over such functions [22]. Content and uncertainty are stored jointly in MemGP: the posterior mean of the distribution specifies expected memory strength at each combination, and the posterior variance specifies uncertainty about that strength. The kernel of the GP prior determines how evidence about one combination informs predictions about others. To give MemGP a finite representation, we create a sparse approximation of this distribution at a finite set of points on feature space, called inducing points [23, 24]. The inducing points are the resource of working memory in MemGP. This resource is discrete, but the precision it supports is continuous, and its allocation among items follows from computing the posterior, without a separate allocation rule [25]. Encoding, maintenance and retrieval are operations on this posterior: encoding infers it from sensory evidence, maintenance updates it from the model's own predictions, and retrieval samples it and divisively normalizes [26, 27] the sampled strengths by those of the other remembered items. Items are localized peaks on a shared representation, so interference [28], competition [7] and bias [29, 30] arise from the representation itself without additional rules.

We instantiate MemGP in a continuous-reproduction task that requires observers to report an object's color at a cued location [2, 3, 31]. The simulation experiments target several classic behavioral effects, including load-dependent precision loss [1–3], retrospective cue benefits [32, 33], and inter-item attraction and repulsion [30, 34]. These experiments place complementary demands on the same representation: increasing load tests shared capacity, retrospective cues test selective maintenance, and inter-item biases test how the maintained information is read out. MemGP reproduces all of these effects and provides an interpretation of human behavioral data. Each behavioral effect is attributed to a component of the posterior: precision falls with load because all items share the same inducing points, attention acts mainly on the uncertainty of the cued memory, and the direction of inter-item bias follows from how the sampled posterior is normalized.

A recurrent circuit derived from the objectives that MemGP optimizes gives the same components a neural interpretation: the posterior mean corresponds to population activity, the posterior covariance to a rapidly updated state, and the prior to the Mexican-hat connectivity that attractor models assume [6, 35, 36].

MemGP also makes testable behavioral and neural predictions. Because one representation predicts recall errors and specifies the dynamics that the circuit implements, the model predicts how a manipulation that changes behavior should change the neural state. A retrospective cue, for example, is predicted to reduce the uncertainty of the cued memory without a comparable change in its expected strength, so the behavioral benefit should be accompanied by a reduction in uncertainty that confidence judgments and neural measurements can test [21].

2 A finite probabilistic representation of working memory

In the continuous-reproduction task (Fig. 2a), an array of N colored items appears briefly. After a blank delay, a location probe identifies an item whose color must be reported on a continuous wheel. MemGP specifies how perceptual evidence is converted into a memory, how that memory changes during the delay, and how it generates a report (Fig. 1d). Encoding forms a posterior over memory-strength functions from sensory evidence, maintenance rehearses that posterior without new sensory input, and retrieval samples the posterior to generate a color report.

2.1 Memory strength and uncertainty

Memory strength is a latent function $f : [-180, 180)^2 \to \mathbb{R}$ on the circular location–color torus (Fig. 1a). Each input is a location–color pair $\mathbf{x} = (\ell, c)$; the value $f(\mathbf{x})$ is a memory-strength field: how strongly the system holds that an item of color c appeared at location ℓ. We place a GP prior over f with a product of periodic kernels matching each dimension's circular geometry:

$$f \sim \mathcal{GP}\big(m(\mathbf{x}),\ k(\mathbf{x}, \mathbf{x}')\big) \tag{1}$$

where $m(\mathbf{x})$ is a constant-valued mean function, $k(\mathbf{x}, \mathbf{x}')$ is a periodic kernel with period 360° and separate lengthscale parameters for location, λ_l, and color, λ_c. Under this kernel, nearby location–color combinations have more strongly correlated memory strengths than distant combinations. Evidence about an observed color at one location can therefore inform predicted strength at nearby colors and locations. The memory representation is the GP posterior over possible strength functions (Fig. 1a), which stores both the content and its uncertainty. At each location–color combination, the posterior mean specifies expected strength and the posterior variance specifies uncertainty about that strength.

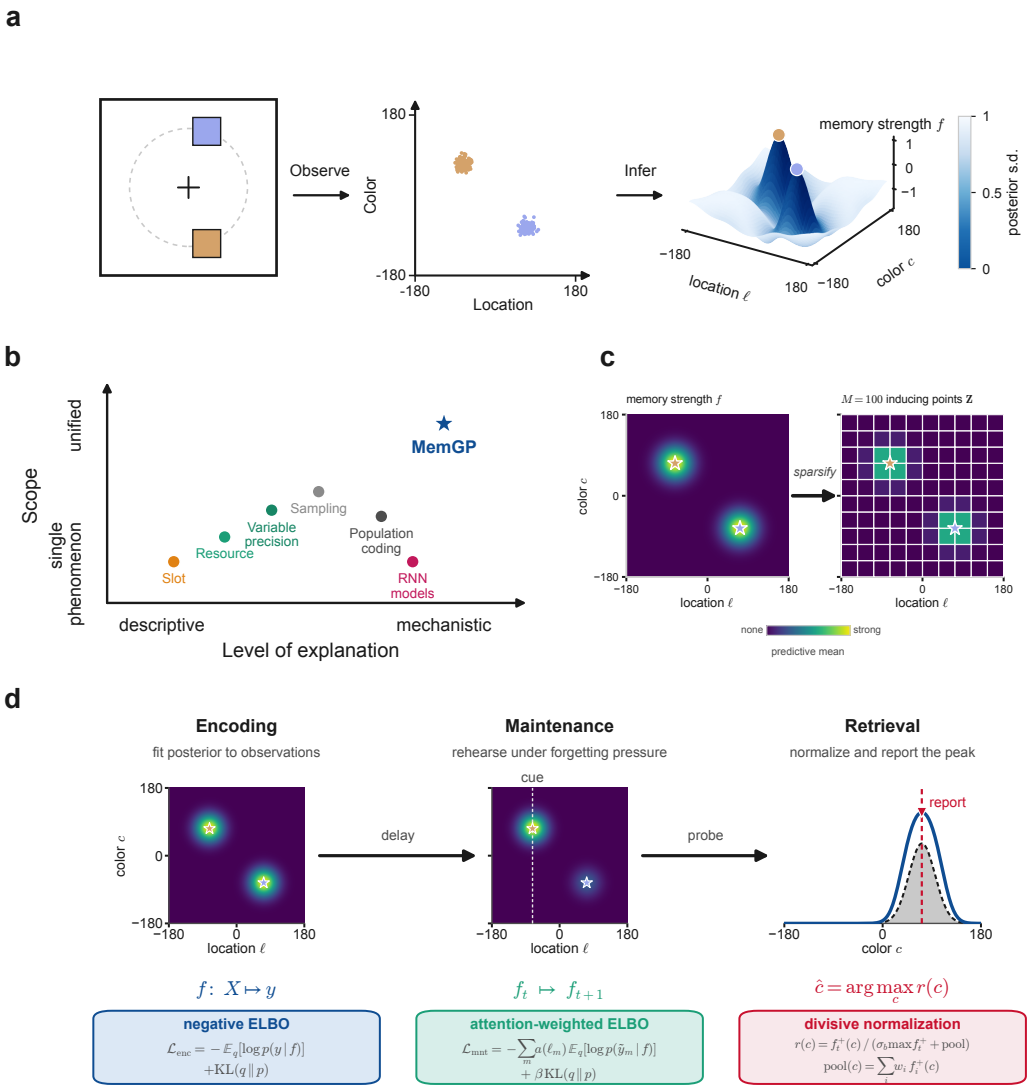


Figure 1: The MemGP framework. **(a)** Encoding as inference. Two colored items provide noisy location–color observations. The GP posterior over the location × color torus has a mean (surface height), with a localized peak at each item, and uncertainty (color indicates posterior standard deviation). **(b)** Schematic comparison of working-memory accounts by scope (single phenomenon → unified) and level of explanation (descriptive → mechanistic). MemGP is shown alongside slot, resource, variable-precision, sampling, population-coding and RNN models; positions are conceptual, not quantitative model-comparison results. **(c)** A finite inducing representation supports a continuous function. The schematic shows memory strength across the feature space and at $M = 100$ inducing points. The regular grid illustrates initialization; inducing positions are learned in the behavioral simulations. Color indicates predictive mean. **(d)** Encoding infers the posterior from perceptual evidence; maintenance uses the preceding posterior mean as rehearsal targets, with selective attention increasing the likelihood weights near the cued location and KL regularization penalizing departure from the prior; retrieval draws a joint posterior sample, divisively normalizes the probed response by the pool of all items, and reports $\arg\max_c r(c)$. Grey fill: normalization pool; black curve: probed response f_t^+; blue curve: normalized readout $r(c)$. The stage illustrations are schematic; full expressions are given in Eqs. 4–5 and Supplementary S1.5.

2.2 Inducing points as a shared computational resource

MemGP approximates the posterior over memory-strength functions using M inducing points, $\mathbf{Z} = \{(\ell_m^{(z)}, c_m^{(z)})\}_{m=1}^M$, and their function values $\mathbf{u} = f(\mathbf{Z})$. The points begin on a regular grid and can move during both encoding and maintenance (Fig. 1c). A mean vector $\mathbf{m} \in \mathbb{R}^M$ specifies expected memory strength at the inducing points. A covariance matrix $\mathbf{S} \in \mathbb{R}^{M \times M}$ specifies uncertainty and dependencies among these strengths. Together they define the Gaussian variational distribution

$$q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, \mathbf{S}). \tag{2}$$

For a queried location–color combination $\mathbf{x}^*$, the GP conditional specifies how its strength depends on the inducing values. Integrating over the uncertainty in those values gives the approximate posterior:

$$q(f(\mathbf{x}^*)) = \int p(f(\mathbf{x}^*) \mid \mathbf{u}) q(\mathbf{u}) d\mathbf{u}. \tag{3}$$

The inducing points provide a finite resource shared by all items. Adding another item does not add inducing points, so the model must represent more location–color combinations using the same number of inducing values. Kernel interpolation predicts memory strength at feature combinations between inducing points, so remembered colors and locations need not coincide with the inducing locations. An item generally depends on several points, and each point can contribute to several items. The computational resource of MemGP is thus discrete while the precision of each item varies continuously.

2.3 Encoding, maintenance and retrieval

During encoding, a procedure that approximates perceptual input supplies noisy location–color observations $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ and evidence strengths $\mathbf{y} = \{y_i\}_{i=1}^n$, with larger strengths for observations closer to an item's location and color (See Methods). A Gaussian likelihood $p(y_i \mid f_i)$ relates this evidence to the latent strength $f_i = f(\mathbf{x}_i)$. To infer the latent strengths from these observations, variational inference adjusts the approximate posterior by optimizing an objective called the evidence lower bound (ELBO). During encoding, MemGP minimizes its negative:

$$\mathcal{L}_{\text{enc}} = -\sum_{i=1}^{n} \mathbb{E}_{q(f_i)}[\log p(y_i \mid f_i)] + \mathrm{KL}\big(q(\mathbf{u}) \,\|\, p(\mathbf{u})\big). \tag{4}$$

The expected log-likelihood term rewards a posterior that predicts the observed evidence strengths. The Kullback–Leibler (KL) divergence measures how far the posterior over inducing values departs from their GP prior. Minimizing the negative ELBO therefore balances agreement with the observations against a penalty for departing from the prior. Because the posterior uses a fixed number of inducing points, all items must be represented within the same finite approximation.

During the delay, external observations are replaced by the model's own previous predictions. The model minimizes:

$$\mathcal{L}_{\text{maint}} = -\frac{1}{M}\sum_{m=1}^{M} a(\ell_m^{(z)})\,\mathbb{E}_q[\log p(\tilde{y}_m \mid f(\mathbf{z}_m))] + \beta\,\mathrm{KL}\big(q(\mathbf{u})\,\|\,p(\mathbf{u})\big) \tag{5}$$

where $\mathbf{z}_m = (\ell_m^{(z)}, c_m^{(z)})$ is the position of inducing point m at epoch t, and $\tilde{y}_m = \mathbb{E}_{q(f)}[f(\mathbf{z}_m)]\big|_{t-1}$ is the posterior mean strength at that point from the previous epoch $t-1$, which serves as the rehearsal target. The rehearsal term favors predictions that remain close to the preceding epoch's memory strengths. The KL term mathematically grounds forgetting over time, penalizing departure from the GP prior with relative weight β (as in [37]. The attention weight $a(\ell_m^{(z)})$ modulates the memory precision near the region $\ell_m^{(z)}$ (Eq. 7). In the continuous-reproduction task, attentional cues are the spatial location $\check{\ell}_j$ of a target item j, and the weights are a Gaussian spatial gain (Fig. 3b) depending only on the distance between $\check{\ell}_j$ and the inducing point's location $\ell_m^{(z)}$. When no cue is presented, all weights equal 1.

At retrieval, MemGP draws one joint posterior sample of memory strength across colors at all occupied item locations. The sampled strength functions are first made nonnegative by subtracting their global minimum. A spatial divisive-normalization rule [26, 27] then divides the nonnegative strength function at the probed location by a baseline term plus a weighted pool of strength functions across remembered locations (Supplementary Eq. S7). Items spatially closer to the probe contribute more to the pool (Supplementary Eq. S6). The model reports the color with the largest normalized response, $\hat{c} = \arg\max_c r(c)$ (Supplementary S1.5, Eq. S8). Thus, encoding and maintenance determine the posterior over strength functions, while sampling and normalization convert that posterior into a color report.

3 Results

3.1 Shared capacity produces load-dependent errors

We first tested whether increasing the number of remembered items reduces recall precision when all items share the same inducing points. Across set sizes $N \in \{1, 2, 4, 6, 8\}$, MemGP reproduced canonical signatures of human continuous-report performance (Fig. 2b–i). As load increased, the error distributions broadened (Fig. 2b,f) and their circular standard deviation (SD) increased (Fig. 2c,g; [2, 3, 31]). Precision, defined here as inverse circular SD, declined approximately as a power law, $1/\text{SD} \propto N^{-\alpha}$ (Fig. 2d,h; [3]). Reports also clustered above chance around the colors of other items in the display, producing swap errors in which a non-target color is reported in response to the target location (Fig. 2e,i; [3]).

MemGP reproduced these patterns without assumptions about the form of the error distribution. Descriptive models [2–5] specify the error distribution directly, for example as a mixture of von Mises components centered on the target and non-target colors and a uniform component [3], and they leave the mechanism that produces the errors unspecified. In MemGP, the error distribution is an outcome of inference, sampling and normalization. The hyperparameters change the magnitude

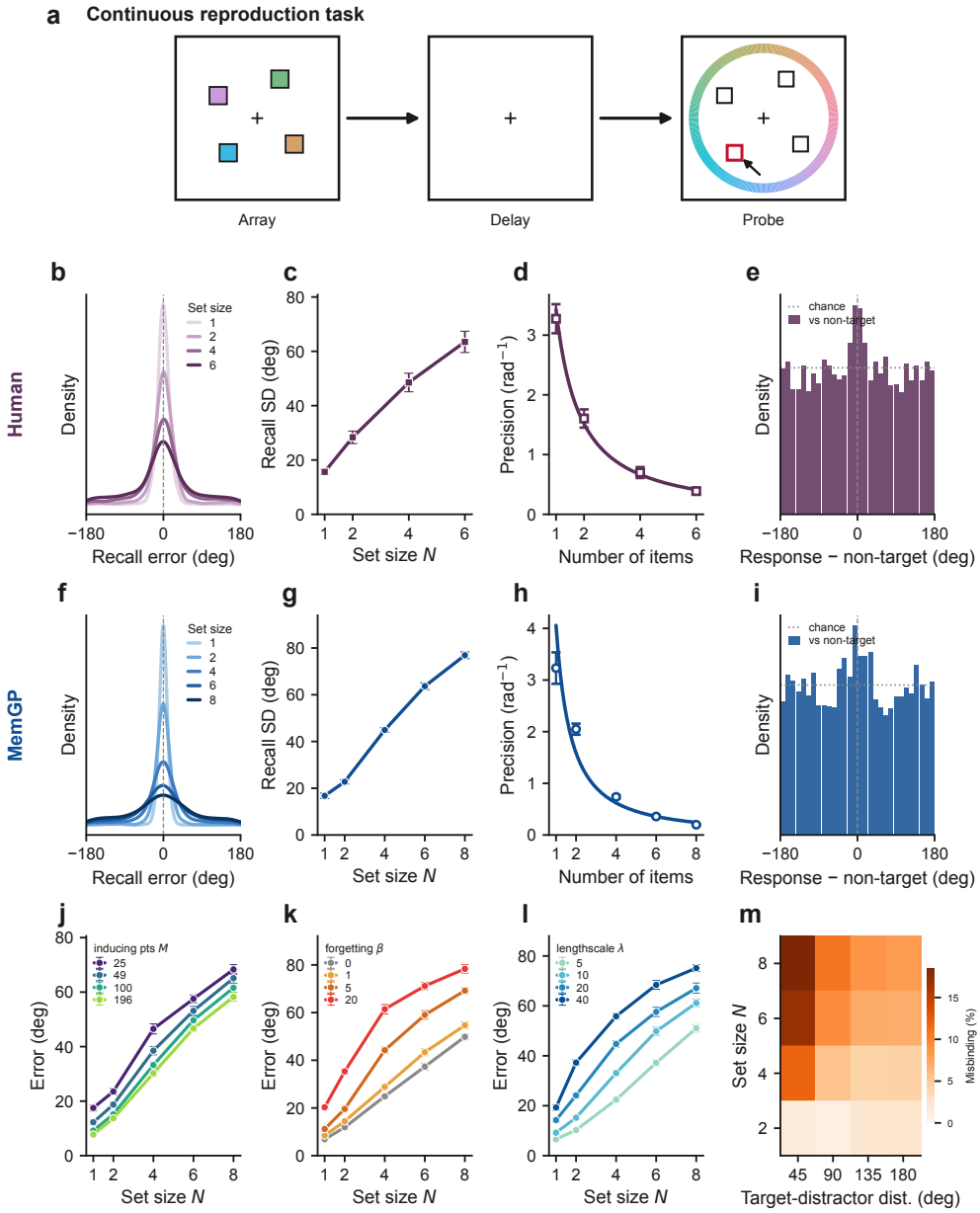


Figure 2: Capacity limits emerge from a shared finite representation. **(a)** Continuous-reproduction task: stimulus array, blank delay, and color-wheel probe at one location. **(b–e)** Human benchmark behavior. **(b,c)** Color error distributions and circular SD vs. set size, as in Ma et al. [5](underlying trials from Bays et al. [3]). **(d)** Precision (inverse circular SD) vs. set size with power-law fit, from Bays et al. [3]. **(e)** Distribution of reported-minus-non-target color differences at $N=4$, from Bays et al. [3]. **(f–i)** MemGP error distributions, circular SD, precision and non-target-relative reports, in the same order as the human panels. **(j–l)** Error vs. set size across inducing-grid sizes $M \in \{25, 49, 100, 196\}$ (j), KL weights $\beta \in \{0, 1, 5, 20\}$ (k), and initial kernel lengthscales $\lambda \in \{5, 10, 20, 40\}$ with location and color kernels set equal (l). Each changes recall error magnitude and preserves the monotonic increase in recall error with set size. **(m)** Misbinding rate (probability that the probe is reported as a non-target item) as a joint function of set size and target–distractor spatial distance: swap errors fall with distance and rise with load. In **(g, h, j–l)**, points are means ± s.e.m. across 20 simulated subjects, with one item probed per trial (100 trials per set size in g,h; 50 in j–l). In **(m)**, each cell is the misbinding rate averaged over all 20 subjects.

of recall error but not its monotonic increase with set size: a larger inducing grid M lowered error at every set size (Fig. 2j), and a larger KL weight β (Fig. 2k) or a wider kernel lengthscale λ (Fig. 2l) raised it at every set size. Set-size effects in MemGP therefore follow from items competing for a finite set of inducing points under the inference process that governs encoding and maintenance.

Because MemGP specifies a mechanism, it can be used to examine interactions that descriptive models do not address, such as the joint dependence of swap errors on set size and target–distractor similarity. In simulations that varied both factors, the probability of misbinding the probe to a non-target decreased with target–distractor distance in the cue dimension, here spatial location, and increased with set size (Fig. 2m; Supplementary S1.6), consistent with human reports [3, 38, 39]. A graded effect of target–distractor similarity is a signature of feature-binding errors [40–42], which population-coding models produce through dedicated binding mechanisms such as conjunction-selective units [41]. MemGP requires no separate binding rule. The kernel correlates memory strengths at nearby locations, so evidence for a neighboring item's color enters the strength profile at the probed location and can determine the reported color.

Encoding precision in human working memory is thought to vary across items and trials [4, 43]. The variable-precision (VP) model captures this variability by assuming that precision follows a Gamma distribution across items and trials, and it fits human data better than models with fixed precision [4]. However, the VP model does not specify why precision varies. MemGP has no parameter for precision variability, yet precision variability arises from posterior uncertainty. Following van den Berg et al. [4], we fitted uniform–von Mises mixtures [2] to MemGP's errors at each set size. We then compared the recovered mixture weight w and circular SD with the predictions of the item-limit, slots-plus-averaging, equal-precision and VP models, each fitted to the same errors (Supplementary Fig. S2). The VP model had the lowest Akaike information criterion (AIC) and Bayesian information criterion (BIC) in all 20 simulated subjects (Supplementary Table S2) and the lowest error on both mixture statistics ($\mathrm{RMSE}_w = 0.092$; $\mathrm{RMSE}_{\mathrm{CSD}} = 0.141$). The trial-by-trial and item-by-item variability in MemGP's precision therefore has the form that the VP model assumes. Posterior uncertainty under capacity-limited inference is thus a candidate mechanism for the precision variability that the VP model treats as a free parameter.

3.2 Selective maintenance reduces posterior variance and improves recall

Working memory is sensitive to retrospective attentional cues (retrocues) presented during the retention interval [32, 33]. Retrocue tasks (Fig. 3a) test how a memory changes during the delay, a process that descriptive models of recall statistics do not address [44]. In MemGP, a retrospective cue increases the likelihood weights of inducing points near the cued item in the maintenance objective (Eq. 7; Fig. 3b). The increased weights give memory strengths near the cued location greater influence on rehearsal. Neutral trials retain equal weights at all inducing points. This spatially reweighted maintenance in MemGP reproduced human performance patterns: lower recall error for cued items, a larger benefit at higher load, and a transfer of the benefit when a second cue selected a different item (Fig. 3c–e; [45–47]).

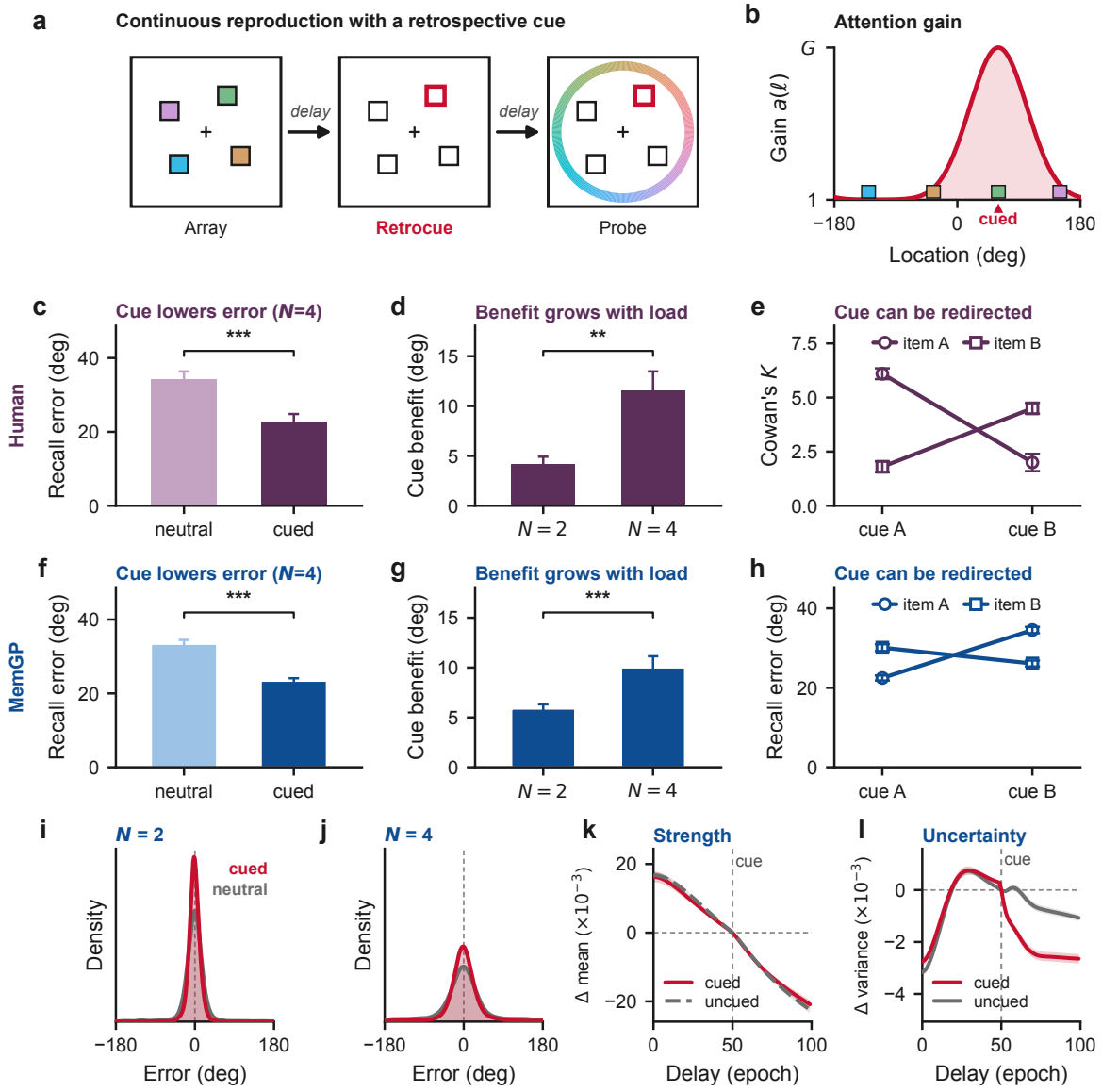


Figure 3: Attention is a reweighting of the maintenance objective. **(a)** Continuous reproduction with a retrospective cue: array, delay, a retrocue indicating one item, delay, and probe. **(b)** Spatial attention gain $a(\ell)$: a Gaussian gain profile centered on the cued location, equal to G at peak and to 1 away from the cue (Eq. 7). Squares on the location axis mark the locations of the four items. **(c–d)** Human continuous-report data from Souza et al. [45]: lower absolute error for focused than neutral items at $N = 4$ (c); cue benefit (neutral minus cued absolute recall error) increasing with set size (d). **(e)** Human cue-redirect pattern from van Moorselaar et al. [46], Exp. 2: Cowan's K for items A and B after an initial cue to A and after a mid-delay redirect to B (change-detection capacity; digitized from their Fig. 2b). **(f–h)** MemGP. Cued error is below neutral, and the benefit is larger at $N = 4$ than at $N = 2$ (f–g). After a mid-delay redirect, error rises for item A and falls for item B (h). **(i–j)** Recall-error distribution, neutral vs. cued, at $N = 2$ (i) and $N = 4$ (j): cueing concentrates the recall-error distribution more strongly around zero, with a larger effect at higher load. **(k)** Change in cued vs. uncued memory strength over the delay (cue at the vertical dashed line): posterior mean declines in both conditions, and the difference between their changes is not significant. **(l)** Change in posterior variance over the same delay, cued item versus the other items: the decrease is larger for the cued item. Bars and points in **(c–h)** are means ± s.e.m. In **(f–h)** there are 20 simulated subjects, 100 trials per condition. Traces in **(k–l)** are means ± s.e.m. across 100 trials.

Load-dependent benefit. In MemGP, cued items were recalled more accurately than neutral items at both set sizes (Fig. 3f). At $N = 2$, mean absolute error was 9.2° for cued items and 15.1° for neutral items, a benefit of $5.9° \pm 0.4°$ (paired $t(19) = 13.39$, $p < 0.001$). At $N = 4$, the corresponding errors were 23.4° and 33.4°, a benefit of $10.0° \pm 1.1°$ (paired $t(19) = 8.82$, $p < 0.001$). The benefit was 4.1° larger at $N = 4$ than at $N = 2$ ($t(38) = 3.40$, $p < 0.001$, one-tailed; Fig. 3g). Selective maintenance in MemGP therefore had a greater effect when more items shared the representation, consistent with load-dependent cue benefits in humans [45, 47].

Flexible reallocation. When a second cue redirected attention from item A to item B during the delay, MemGP's error rose by 12.1° for A and fell by 4.0° for B (Fig. 3h). The difference between the error changes for the two items, $\Delta A - \Delta B$, was 16.1° (95% CI $[13.3, 18.8]$; paired $t(19) = 12.24$, $p < 0.001$). The loss of precision for the previously cued item and improvement for the newly cued item agree in direction with the double-cue results reported by van Moorselaar et al. [46] (Fig. 3e). Their experiment measured change-detection capacity, whereas the MemGP simulations measured continuous-report error, so the comparison concerns the direction of the effect across different behavioral measures (Cowan's K vs. recall error).

Maintenance dynamics. Cueing concentrated the recall-error distribution more strongly around zero, with a larger change at $N = 4$ (Fig. 3i–j). To relate this behavioral benefit to changes in the maintained representation, we examined the posterior mean and variance separately. The posterior mean, which measures expected memory strength, declined for both cued and uncued items; the difference in their changes from cue onset to delay end was not significant (paired $t(99) = 1.41$, $p = 0.16$; Fig. 3k). In contrast, posterior variance, which measures uncertainty about memory strength, decreased more for the cued item than for the mean of the other items (paired $t(99) = -7.53$, $p < 0.001$; Fig. 3l). This dissociation explains why the model needs both strength and uncertainty: lower recall error can accompany reduced posterior variance even while mean strength declines (Supplementary S3.3).

3.3 Shared representations and normalization produce attraction and repulsion

When several items are held in memory, the recalled feature of one item is often biased by its neighbors: the report may shift toward a neighboring feature (attraction) [48] or away from it (repulsion) [29, 49]. These biases have been interpreted as adaptive [30]. Repulsion can keep a few confusable items distinct, and attraction can stabilize noisy estimates of individual items with information from the ensemble [34, 48, 50, 51]. Mechanistic models produce attraction [28, 52] and, in some cases, repulsion [8, 53], but they mainly address how the bias depends on the feature distance between items. In human data, the direction and size of the bias also depend on multiple confounding factors [30]. MemGP accounts for these dependencies through its retrieval rule (Fig. 4a). The sampled strength profile over color at the probed location is divided by a baseline term plus a weighted pool of the sampled profiles of all items in the display. The direction of the bias depends on where the pool lies relative to the target profile. When a neighbor's color is close to the target's, kernel generalization shifts the target profile toward the neighbor, and the pool, formed from the

same overlapping profiles, has nearly the same shape over color as the target profile. Division by the pool therefore leaves the peak position largely unchanged, the normalized response follows the shifted profile, and the report is attracted (Fig. 4a, left). When the neighbor's color is farther away, the pool is larger on the side of the target profile that faces the neighbor, so division suppresses that side and the report is repelled (Fig. 4a, right).

We quantified the direction of bias using the model-free measure of Chunharas et al. [30], which counts reports rather than fitting a distribution to the errors:

$$\text{bias} = 100 \times \frac{n_{\text{away}}}{n_{\text{total}}} - 50, \tag{6}$$

where n_{away} counts reports whose error points away from the non-target color, and n_{total} counts all reports. Subtracting 50 centers the measure on equal proportions of errors toward and away from the non-target. Positive values indicate repulsion and negative values indicate attraction; values are expressed in percentage points relative to 50%.

At low load ($N = 2$), bias was near zero at the smallest color separation ($2.4 \pm 1.0\%$ at $20°$) and increased to repulsion, peaking around 75–105° ($23.8 \pm 1.1\%$ at $105°$; means $\pm$ s.e.m. across 20 simulated subjects; Fig. 4b). At high load ($N = 6$), with distractor colors placed on the same side of the target, clustered separations produced attraction ($-27.1 \pm 0.9\%$ at $20°$). Repulsion returned at wider separations ($7.2 \pm 1.3\%$ at $105°$; Fig. 4c). The same posterior-sampling and normalization rule therefore generated attraction or repulsion depending on memory load and the arrangement of colors, consistent with the human tendency toward attraction under high load [30].

We also tested whether retrospective attention and the noise in encoding observations change the direction or magnitude of recall bias. Cueing the target during the delay increased repulsion from $16.7 \pm 1.1\%$ to $21.1 \pm 1.1\%$ (Fig. 4d). Noisier encoding (color sampling noise $\sigma_c = 60°$) also produced greater repulsion than more precise encoding ($\sigma_c = 10°$; $35.4 \pm 0.7\%$ versus $13.9 \pm 1.4\%$; Fig. 4e), reproducing the direction of the encoding manipulation in Chunharas et al. [30] and Lively et al. [34].

In the human data, lower circular SD was associated with greater repulsion (Pearson $r = -0.81$, 95% CI $[-0.91, -0.60]$; Fig. 4f). We found the same strong negative correlation in the MemGP simulations (Pearson $r = -0.88$, 95% CI $[-0.95, -0.75]$; Fig. 4g). For both human participants and MemGP simulations, repulsion biases are thus strongest when recall is more precise.

In summary, attraction and repulsion in MemGP arise from one retrieval rule applied to a shared memory representation. Feature distance, load, attention, encoding noise and recall precision each change the direction or size of the bias, as they do in human reports.

3.4 Variational inference implemented by recurrent circuit dynamics

MemGP accounts for the behavioral effects of memory load, retrospective attention and inter-item interactions. To relate the computations that produce these effects to a circuit, we interpret the mean as population activity, with each unit corresponding to a population tuned to a region of

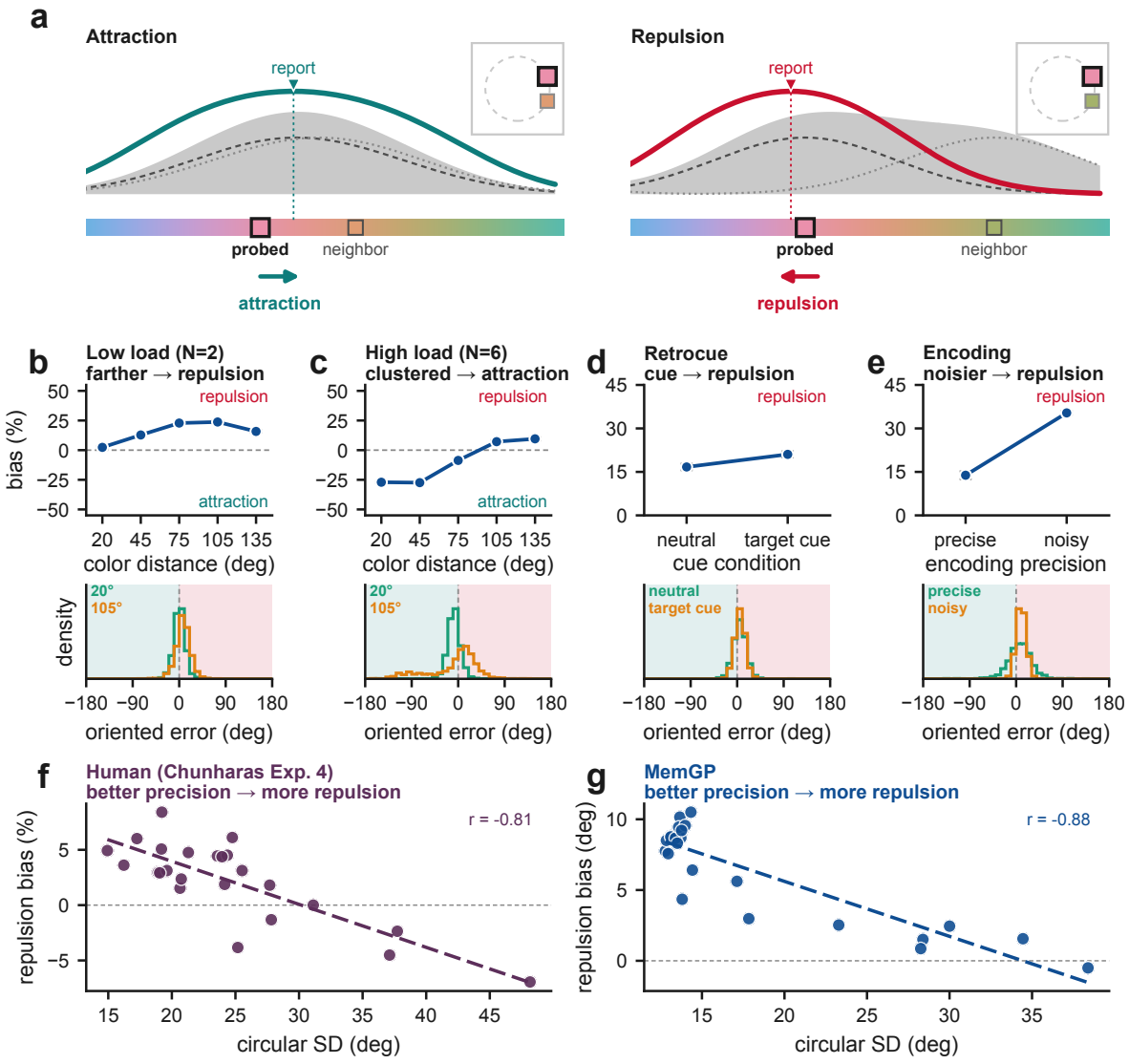


Figure 4: Inter-item attraction and repulsion depend on memory demands. **(a)** Two scenarios for the color profile at the probed location (schematic). Left: a neighbor with a similar color shifts the target profile toward the neighbor; the pool has nearly the same shape, and the readout follows that shift (teal). Right: a neighbor with a more distant color contributes an asymmetric normalization pool that shifts the readout away from the neighbor (red). Dashed curve: target profile; dotted curve: neighbor profile; grey fill: pool; bold: normalized readout. **(b,c)** Bias vs. color distance at low load ($N = 2$, b) and high load ($N = 6$, c); histograms in both panels highlight 20° and 105°. **(d)** A retrocue on the target modestly increases repulsion. **(e)** At color distance 135°, noisier encoding ($\sigma_c = 60°$) increases repulsion relative to precise encoding ($\sigma_c = 10°$), matching the Exp. 4 signature in Chunharas et al. [30]. In **(b–e)**, points are means ± s.e.m. across 20 simulated subjects, 100 trials per condition (s.e.m. between 0.8% and 1.4%, nearly invisible). Histograms show oriented-error distributions (left/right shading: attraction/repulsion). **(f, g)** Lower circular SD predicts stronger repulsion in humans and MemGP. **(f)** is reproduced from Chunharas et al. [30] Exp. 4 data (24 subjects). In **(g)**, each MemGP simulated subject ($n = 24$) is tested at target–distractor color distances $\{20, 40, 60, 80, 100, 120, 140\}°$; circular SD and signed angular bias are averages across those distances.

location–color space, and the covariance as a separate, rapidly updated state. The variational objectives then specify how both states change (Fig. 5a). With the inducing locations, prior mean, kernel parameters, and likelihood noise fixed, the Gaussian variational objective is quadratic in the mean (Supplementary Eqs. S16 and S21). Moving the mean down the gradient of this objective gives a linear update consisting of recurrent interactions and an input drive (Supplementary Eqs. S17 and S22).

For maintenance, the resulting dynamics are $\tau\dot{\mathbf{m}} = -\mathbf{W}\mathbf{m} + \mathbf{b}$ (Supplementary Eq. S23), where $\mathbf{W} = \mathbf{A}/(\sigma^2 M) + \beta\mathbf{K}_{ZZ}^{-1}$ is the recurrent operator. The diagonal matrix $\mathbf{A}$ contains the attention gains, σ^2 is the likelihood noise variance, M is the number of inducing points, and β is the maintenance KL weight. The matrix $\mathbf{K}_{ZZ}$ is the prior covariance of the inducing values, so its inverse is their prior precision. The input drive $\mathbf{b}$ depends on the current rehearsal target and prior mean; τ sets the timescale of the ideal gradient dynamics (Supplementary S3).

The prior-precision term in $\mathbf{W}$ supplies interactions between units that remain fixed when inducing locations and kernel parameters are fixed. Because the activity equation contains $-\mathbf{W}\mathbf{m}$, the prior-dependent interactions have sign $-\mathbf{K}_{ZZ}^{-1}$. For the periodic prior and grid shown in Fig. 5b, the off-diagonal weights exhibit local excitation and surrounding inhibition, the Mexican-hat pattern used in working-memory attractor models. Thus, for this prior and inducing grid, the Mexican-hat connectivity is obtained by differentiating the inference objective.

Updating mean activity alone would leave the represented uncertainty unchanged. Differentiating the same variational objective with respect to covariance gives the additional update, $\tau\dot{\mathbf{S}} = -(\mathbf{W} - \beta\mathbf{S}^{-1})$, with stationary covariance $\mathbf{S}^* = \beta\mathbf{W}^{-1}$ (Supplementary Eq. S26). The recurrent operator determines how mean activities interact. The covariance update determines uncertainty and dependencies among the strengths sampled at retrieval. $\mathbf{S}$ therefore can be interpreted as fast weights, relating this formulation to architectures that distinguish slow and fast connections [54–56].

The circuit derivation fixes inducing locations and kernel parameters, whereas MemGP still updates them during maintenance. To assess how this difference affects recall, we compared the recurrent circuit with MemGP on matched trials. For each matched trial, the recurrent circuit and MemGP began from the same posterior obtained after encoding. During maintenance, the recurrent circuit fixed the inducing locations, kernel and prior mean at their encoding-end values, while MemGP continued to update them. Mean absolute recall error, estimated from 100 reports per trial, agreed closely between the circuit and MemGP (Pearson $r = 0.987$, 95% CI $[0.983, 0.990]$; 200 trials pooled across neutral and retrocue conditions; Fig. 5c).

We next measured how the represented color changed during circuit maintenance. At each epoch, we reconstructed the posterior mean over color at the remembered location and tracked the position of its peak. The peak moved farther from its delay-onset position as maintenance progressed (Spearman $\rho = 0.56$, trial-level bootstrap 95% CI $[0.50, 0.62]$; mean displacement $0.02° \to 1.26°$ over 100 epochs; 100 trials; Fig. 5d). The displacement resembles the delay-dependent drift associated with recall errors in prefrontal recordings [36], although the present simulation does not fit those recordings or specify the same noise mechanism.

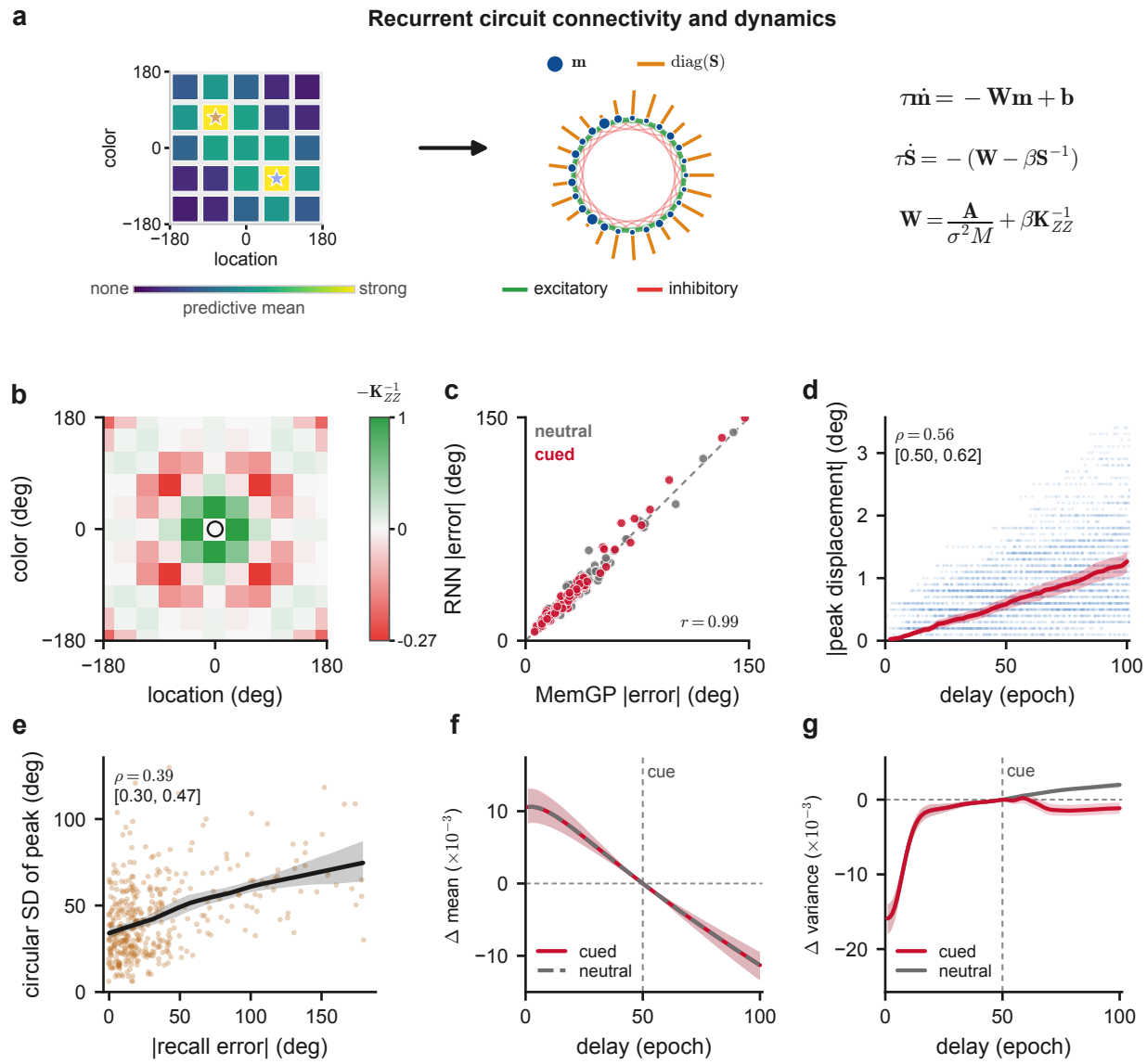


Figure 5: Variational updates yield recurrent dynamics with center–surround interactions. (a) A schematic inducing grid ($M = 25$; square color is the predictive mean) storing two items is transformed as 25 recurrent units on a ring. Circle size indicates expected strength $\mathbf{m}$; spoke length indicates marginal variance $\mathrm{diag}(\mathbf{S})$. Green and red arcs indicate excitation and inhibition. **(b)** Outgoing weights $-\mathbf{K}_{ZZ}^{-1}$ from one inducing unit (the open circle; self-coupling removed) on the 10×10 grid ($M = 100$), computed as $-(\mathbf{K}_{ZZ} + \varepsilon\mathbf{I})^{-1}$ with jitter $\varepsilon = 10^{-3}$ (Supplementary Fig. S3). Weights are scaled by their largest absolute entry; green and red indicate excitation and inhibition. **(c)** Recall error of MemGP and the recurrent circuit on matched trials. Each point is one trial's mean absolute error over 100 posterior-sample reports, in the neutral (grey) and retrocue (red) conditions. Dashed line: identity. Pearson $r = 0.99$ across 200 trials. **(d)** Absolute displacement of the posterior-mean peak from its position at delay onset in the recurrent circuit. Displacement grows with delay under KL regularization (100 trials; Spearman $\rho = 0.56$). **(e)** Circular SD (over 200 posterior-sample reports) of the posterior-mean peak against the absolute error of an independent report on the recurrent circuit (100 trials at $N = 4$; Spearman $\rho = 0.39$). **(f, g)** Change in posterior mean (f) and variance (g) at the probed item on the recurrent circuit, from their values at cue onset. Each trial was encoded once and then maintained twice, with a retrocue and with a neutral cue ($N = 4$, 100 trials). Lines in **(d, f, g)** are across-trial means; in **(e)**, points are items and the line is a local-linear fit. Bands are bootstrap 95% intervals.

To test whether uncertainty in the circuit's reports predicts error, we generated one report per item and estimated report variability from separate samples. The circular SD of 200 additional sampled reports was associated with absolute recall error (Spearman $\rho = 0.39$, bootstrap 95% CI $[0.30, 0.47]$, $p < 0.001$; 100 trials at $N = 4$; Fig. 5e). This association between report variability and error is consistent with evidence that neural uncertainty estimates relate to behavior in perception and working memory [21, 57].

To test whether the circuit also reproduces the MemGP's selective reduction in uncertainty after a retrocue, we compared posterior-mean and posterior-variance trajectories in cued and neutral delays starting from the same encoded posterior. The conditions produced similar mean trajectories but different variance trajectories: cueing reduced variance at the probed item while neutral maintenance increased it (Fig. 5f–g). The circuit therefore preserves the distinction observed in MemGP: retrospective attention can reduce uncertainty about memory strength without a corresponding increase in expected strength.

4 Discussion

Descriptive and mechanistic models of working memory have lacked a shared representation. MemGP supplies this representation as a posterior over a memory-strength function. The posterior stores content and uncertainty jointly, a finite set of inducing points is the resource that supports it, and encoding, maintenance and retrieval are operations on it. In the simulations, this one posterior reproduced load-dependent errors, retrospective cue benefits, and inter-item attraction and repulsion, and each effect was attributed to a component of the posterior. Precision fell with load because all items share the inducing points; a retrocue reduced the variance of the cued memory more than it changed its mean; and the direction of inter-item bias followed from how the sampled posterior was normalized. A recurrent circuit derived from MemGP has the center–surround connectivity assumed by attractor models and agreed closely with MemGP in recall error. Its dynamics also showed delay-dependent drift of the represented color, an association between report variability and recall error, and a selective reduction in uncertainty after a retrocue. MemGP therefore provides a common algorithmic framework that connects behavioral and circuit accounts of working memory.

4.1 Discrete capacity and continuous precision

MemGP reconciles discrete capacity and continuous precision by assigning them to different properties of the memory representation. The number of inducing points is finite, while their locations and the mean and covariance of their function values vary continuously. All remembered items share these inducing points. Increasing the number of items therefore requires the same finite representation to describe more location–color combinations, reducing recall precision in the simulations. A finite number of inducing points does not imply a fixed number of stored objects: an item can depend on several points, and a point can contribute to several items. The model consequently combines a discrete computational resource with continuously varying precision for individual memories.

Consistent with this distinction, the variable-precision model described MemGP-generated errors better than the item-limit and slots-plus-averaging models, so a finite number of inducing points did not produce the error distributions expected from a fixed item limit.

Previous models have linked discrete capacity and continuous precision through stochastic sampling, psychophysical similarity, and resource-rational allocation [17, 18, 25]. These models mainly describe either how a limited resource is allocated among items or how a report is generated from memory signals. By contrast, MemGP uses one posterior throughout a trial: the posterior is updated from sensory evidence during encoding, selectively maintained under attention, and sampled to generate a color report.

Among these models, the target confusability competition (TCC) model [18] is most directly comparable to MemGP at retrieval: both select the largest of correlated noisy strength signals for candidate colors. The models differ in the source of these correlations. In MemGP, they arise from the GP posterior, whose covariance depends on the kernel, and the sampled signals undergo spatial divisive normalization before selection. In the TCC model, they are constrained by psychophysical similarity measurements. Incorporating measured perceptual similarity into the GP kernel would allow a more direct comparison of how the models predict recall errors. Quantitative fits to human data would then be needed to distinguish their predictions.

4.2 Variational inference under computational constraints

In MemGP, variational inference gives encoding and maintenance the same form of optimization: preserve information while penalizing departure from a prior. During encoding, the likelihood term rewards agreement with sensory observations. During maintenance, it rewards agreement with memory strengths predicted at the preceding epoch. In both stages, the KL term penalizes departure of the inducing-value posterior from the GP prior, and the number of inducing points limits the approximation used to represent that posterior. Attention increases the maintenance likelihood weights near a cued location, giving the corresponding rehearsal targets greater influence on the update. This explicit trade-off between fidelity and regularization relates MemGP to resource-rational theories of memory and to accounts of attention as precision control [25, 58].

Because attention reweights a maintenance objective that depends on both the posterior mean and covariance, a cue can affect expected memory strength and uncertainty differently. In the retrocue simulations, posterior variance decreased more for the cued item, whereas the difference in posterior-mean changes was not significant. This result identifies a way that attention can improve recall without necessarily strengthening the memory, which distinguishes it from proposals that a retrocue strengthens the cued representation [33]. Experiments combining recall reports, confidence judgments and neural measurements could test whether the behavioral benefit is accompanied by the predicted reduction in uncertainty [21].

The KL penalty also relates MemGP to efforts to derive neural computation from resource and energy constraints [59, 60]. In MemGP, this penalty is an information-theoretic cost on the Gaussian posterior over inducing values and does not directly measure metabolic expenditure. In Poisson

variational models, by contrast, the KL term becomes proportional to the prior firing rates of model neurons, which gives it an interpretation as metabolic cost [61]. A comparable interpretation of MemGP's Gaussian KL penalty would require a biophysical model of how neurons encode and update the posterior mean and covariance.

4.3 Connectivity and units of the recurrent circuit

Recurrent models of persistent activity [6, 36] and Bayesian attractor networks [62] assume connectivity with local excitation and surrounding inhibition, and this Mexican-hat pattern maintains a localized bump of activity during the delay [35, 63]. In MemGP, the same pattern follows from the prior. The GP kernel specifies how memory strengths covary across feature combinations under the prior. In the circuit derivation, this prior covariance determines recurrent interactions through its inverse, the prior precision matrix. For the periodic kernel used here, these interactions excite the nearest units and inhibit a surrounding ring (Fig. 5b). The pattern depends on the smoothness of the kernel and on the conditioning of $\mathbf{K}_{ZZ}$ (Supplementary S3 and Supplementary Fig. S3).

The units of the derived circuit have a candidate neural counterpart. A prefrontal neuron with a memory field shows sustained delay activity when the remembered location lies within a limited region of space, and these regions differ across neurons so that together they cover the visual field [64]. In MemGP, each inducing point could correspond to the neural population whose memory fields share a region of feature space. The activity of the unit would then be the mean firing rate of this population, representing expected memory strength for that region; the rehearsal term sustains this activity during the delay, and the prior-precision term couples populations tuned to neighboring regions. Because the inducing points are defined over features (e.g., location and color) jointly, the closest counterparts are neurons with conjunctive tuning, which have also been reported in prefrontal cortex [65].

4.4 Slow weights, fast weights and activity-silent memory

The derived circuit contains two kinds of coupling that change on different timescales. The prior-dependent part of the recurrent operator is fixed during the delay once inducing locations and kernel parameters are fixed, and the covariance $\mathbf{S}$ is updated at every epoch. These two quantities correspond to the slow and fast weights of architectures that distinguish the two [54, 55]. The mean activity represents expected memory strength, and the covariance represents uncertainty and dependencies among inducing values. The covariance determines the variability and correlations of sampled strengths, which in turn affect the color selected at retrieval. Representing expected strength in activity and uncertainty in a rapidly updated state therefore relates the circuit to both persistent-activity accounts and theories in which short-term synaptic changes contribute to memory [9, 66]. A synaptic implementation of the covariance update remains to be established.

The proposed role of fast synaptic changes raises a further issue: whether memory can remain accessible without persistent activity. Activity-silent theories propose that information can remain accessible when stimulus-specific firing is weak or absent [10, 67]. MemGP assigns a possible role to

fast synaptic changes: they could represent the covariance needed for uncertain recall. The present simulations, however, do not demonstrate storage and recovery of item identity during complete activity silence. The covariance describes uncertainty around the mean; storing the covariance alone need not retain the mean that identifies the remembered items. Demonstrating activity-silent retention would therefore require a mechanism that stores and restores item-specific information when activity ceases and resumes. The model instead motivates tests of how activity and fast synaptic variables jointly determine content and uncertainty, consistent with evidence that persistent and activity-silent processes can interact [68, 69].

4.5 Extensions and limitations

The Gaussian-process formulation separates assumptions about baseline memory strength, feature similarity and computational capacity. The prior mean specifies baseline expected strength. The kernel specifies how strengths at different feature combinations covary and therefore how evidence generalizes across those combinations. The inducing points determine the finite approximation used to represent the posterior. Each component can be changed while retaining the variational inference procedure. Although GP kernels can be defined over other feature spaces, the present behavioral tests are restricted to location and color. Extending to multi-feature binding will require higher-dimensional GPs or structured kernels, in which case deep kernel learning [70] or spectral mixture kernels could allow the covariance structure to be learned from data.

Long-term memory could influence MemGP through the GP prior. Learning across trials could shape the prior mean and kernel to reflect familiar feature combinations and their relationships, consistent with evidence that learned stimulus statistics alter working-memory dynamics [71]. These changes would affect both posterior inference and the recurrent connectivity derived from the prior precision. As the posterior relaxes toward a learned, structured prior during maintenance, recall could become biased toward familiar stimuli. Retaining part of the memory state between trials could also allow the model to test serial dependence, in which a current report depends on previously presented items [69]. Both extensions require learning or state that carries over between trials, which the present simulations do not include.

The same components could describe perception and decision making. Encoding in MemGP is already an inference from noisy sensory observations to a posterior over feature space, and reading out this posterior without a delay would give a model of perceptual estimation that carries its own uncertainty [19, 57, 72]. Because the posterior stores content and uncertainty jointly, a downstream decision can weight a remembered feature by its reliability, and the posterior variance provides a candidate source for confidence judgments [21]. Gaussian-process posteriors over functions have also been used to model how people generalize and explore in reward-based decision tasks [73, 74], so the representation that MemGP uses for memory content could be shared with models of learning and choice.

The present simulations establish qualitative agreement with the tested human behavioral effects, but quantitative fits to individual participants remain to be evaluated. Fitting participant data,

potentially with Bayesian optimization [75], would test whether distinct parameter settings can be distinguished from their behavioral predictions and whether MemGP predicts recall more accurately than alternative models.

The recurrent circuit derived from MemGP has three further limits. The covariance update requires $\mathbf{S}^{-1}$ at every step, and computing or updating this inverse through local synaptic plasticity has no established mechanism. Time is measured in optimization epochs rather than seconds; relating the dynamics to real delay intervals, for example by testing whether maintenance persists across delays of different lengths as prefrontal delay activity does [64], would make the dynamic predictions directly testable. The circuit is also a rate model in which each inducing point is a single unit; the proposed correspondence to a tuned neural population has not been implemented with spiking neurons or with multiple neurons per inducing point.

In sum, MemGP treats working memory as inference over continuous feature spaces, approximated by a finite set of inducing points shared by all items. The posterior mean and covariance determine expected memory strength and uncertainty, attention changes which rehearsal targets receive greater weight, and posterior sampling generates variable recall responses. Deriving recurrent equations for the same mean and covariance connects these algorithms to candidate neural dynamics. Working-memory behavior and its candidate neural circuits thus follow from a single inference process, providing a unified account across descriptive, algorithmic and circuit levels of explanation.

5 Methods

Full mathematical details are provided in the Supplementary Methods. Here we summarize the key components of MemGP.

5.1 Model specification

MemGP implements a sparse variational Gaussian process [23] using GPyTorch [76]. The input space is two-dimensional: each input is a pair $\mathbf{x} = (\ell, c)$, with spatial location $\ell \in [-180°, 180°)$ and color $c \in [-180°, 180°)$, both circular with period 360°. The GP prior mean $m(\mathbf{x})$ is constant across feature space and learned during inference. The prior covariance is a scaled product of two periodic kernels (Eq. 1). The variational posterior over inducing values is parameterized by a mean $\mathbf{m}$ and a covariance $\mathbf{S}$, represented through its Cholesky factor, at $M = m^2$ inducing points $\mathbf{Z} = \{(\ell_m^z, c_m^z)\}_{m=1}^M$ initialized on a regular grid (default $m = 10$, $M = 100$), with inducing values $\mathbf{u} = f(\mathbf{Z})$. The inducing positions are learned during encoding and maintenance. A Gaussian likelihood $p(y_i \mid f_i)$ with fixed, small observation noise completes the generative model (Supplementary Table S1).

During cued maintenance, inducing point m receives a spatial attention weight that depends only on its location coordinate. For a cue at $\check{\ell}_j$,

$$a(\ell_m^z) = 1 + (G-1)\exp\left(-\frac{d(\ell_m^z, \check{\ell}_j)^2}{2\sigma_s^2}\right) \tag{7}$$

with peak gain G, Gaussian spatial width σ_s, and $d(\cdot, \cdot)$ circular distance. When no cue is presented, all weights equal 1.

5.2 Simulation protocol

Stimulus generation. Unless otherwise specified, each trial used N locations sampled without replacement from eight equally spaced positions on a circle, with a common random rotation. Colors were sampled uniformly from $[-180°, 180°)$ and converted to integer degrees. A surrogate perceptual procedure generated $n_{\text{obs}} = 1000$ observations per item, with independent Gaussian location and color deviations ($\sigma_\ell = \sigma_c = 20°$), wrapped onto the circular axes. Each observation was assigned an evidence-strength target given by the product of Gaussian functions of its circular distances from the encoded center. These values entered the Gaussian likelihood as regression targets y_i. The model does not simulate the perceptual computations that would generate this evidence.

Inference. Encoding minimizes the negative ELBO (Eq. 4) using the Adam gradient-based optimizer. Maintenance minimizes the rehearsal objective (Eq. 5), whose targets are the preceding posterior means at the inducing points. Both stages update the inducing locations, kernel hyperparameters, prior mean, variational mean and variational covariance. During cued maintenance, attention reweights the likelihood terms according to Eq. 7. The maintenance optimizer inherits its first- and second-moment estimates from encoding and switches to the maintenance learning rate. Epoch counts, learning rates, and the KL weight β are listed in Supplementary Table S1. Kernel lengthscales in that table specify their initial values.

Retrieval. One joint sample was drawn from the latent GP posterior over 360 evenly spaced color values at every occupied item location, preserving covariance across both colors and items. These samples did not include an additional draw of Gaussian likelihood noise. Subtracting the minimum sampled strength across all locations and colors from every sampled strength made the responses nonnegative. Spatial divisive normalization [26, 27] was then applied, and the color with the maximum normalized response was reported (Supplementary S1.5). The surround width was 60° and the baseline fraction was 0.8 across the set-size, retrocue and bias experiments. Repeated posterior draws were used only where specified to estimate distributions, swap rates or circuit readout summaries.

5.3 Experiments

Set size. Load was varied over $N \in \{1, 2, 4, 6, 8\}$ (20 simulated subjects, 100 trials per set size; uniform attention). As in human continuous-report tasks, one randomly located item was probed per trial, so every set size gave each subject the same number of errors. Circular SD and precision (inverse circular SD) were computed per subject and summarized as the subject mean ± s.e.m. Non-target-relative errors at $N = 4$ were the circular difference between each report and every non-target color (Fig. 2i). Hyperparameter sweeps repeated the set-size series while varying inducing-grid size $M \in \{25, 49, 100, 196\}$, KL weight $\beta \in \{0, 1, 5, 20\}$, or initial kernel lengthscale $\lambda \in \{5, 10, 20, 40\}$ (location and color kernels set equal).

Misbinding. Swap rates as a joint function of set size and target–distractor spatial distance used $N \in \{2, 4, 6, 8\}$ (20 simulated subjects, 30 displays per set size). Every item in each display was probed, with 100 reports generated from independent posterior samples for each probe. A report counted as a swap toward a given non-target if it fell within 30° of that item's color and outside 30° of the target. Pairs were grouped by circular spatial distance $\{45°, 90°, 135°, 180°\}$ (Supplementary S1.6).

Retrocue. Neutral trials used uniform attention throughout a 100-epoch delay. Cued trials applied Eq. 7 from epoch 50, with the Gaussian centered on the probed item's location. Load dependence compared $N = 2$ and $N = 4$ (20 simulated subjects, 100 trials per condition). Cue-flexibility trials ($N = 4$, 200-epoch delay; 20 simulated subjects, 100 trials; Fig. 3h) applied the same gain first to a randomly chosen item A (epochs 50–99) and then to a different item B (epochs 100–199). Absolute recall error for A and B was read out at the end of each cue window.

Attraction/repulsion bias. Displays used a fixed target at location 0°. At $N = 2$, one distractor sat at 90°; at $N = 6$, five distractors occupied the remaining equally spaced locations, all on the same side of the target in color, clustered about the nominal color distance. Color distance was swept over $\{20, 45, 75, 105, 135\}°$ (20 simulated subjects, 100 trials per condition). Cued vs. neutral comparisons and the encoding-noise comparison ($\sigma_c = 10°$ vs. 60°) used $N = 2$ at 135°. For the precision–repulsion correlation (Fig. 4g), each of 24 simulated subjects drew encoding noise from Uniform[10°, 40°] and was tested at color distances $\{20, 40, 60, 80, 100, 120, 140\}°$; circular SD and the mean oriented error (degrees; trials with absolute error above 40° excluded) were averaged across distances. Human subject means in Fig. 4f were recomputed from Chunharas et al. [30] Exp. 4, aggregating over encoding × distance cells. Bias in Fig. 4b–e was the model-free measure in Eq. 6.

Recurrent circuit implementation. Matched-trial recall compared MemGP to the recurrent circuit (200 trials, pooled across neutral and retrocue). For each trial, MemGP and the recurrent circuit begin from the same posterior obtained after encoding. The recurrent circuit fixes the inducing locations, kernel and prior mean at encoding end, while MemGP continues to update them through the delay. The comparison in Fig. 5c uses mean absolute error over 100 sampled reports per trial, with matched random seeds for readout. The recurrent circuit retains the encoding optimizer state and updates the variational mean and Cholesky covariance in GPyTorch's whitened coordinates. Whitening transforms the inducing variables so that their prior has zero mean and identity covariance (Supplementary S3.3).

Bump displacement (Fig. 5d) is the movement of the peak read out from $\mathbf{m}$: the argmax, on a 0.1° color axis, of the posterior mean at the item's location, measured from its position at delay onset ($N = 1$, 100 trials). The argmax was taken within 45° of the peak's position at the previous epoch, starting from the item's color, so that a second peak elsewhere on the color axis is not read out as displacement. On the inducing grid that mean equals $\mathbf{m}$; off the grid it is the kernel interpolation of $\mathbf{m}$ (Supplementary S3.2).

For Fig. 5e, each of four items on each of 100 trials received one sampled report; uncertainty

was the circular SD of 200 additional reports from the same posterior and readout. The circular SD and the independent report error therefore used separate draws.

Trial counts for the paired cued and neutral trajectories are given in Fig. 5. The mean and variance trajectories were obtained by running cued and neutral maintenance from the same encoded posterior ($N = 4$, 100 trials), with the cue applied from epoch 50, and are plotted as the change from each trial's value at cue onset.

5.4 Model comparison

The comparison fitted the benchmark models of van den Berg et al. [4] to MemGP's simulated errors to assess which descriptive model best summarized those errors. The fitted models consist of Item-Limit (IL), Equal-Precision (EP), Slots-plus-Averaging (SA), and Variable-Precision (VP), which were fitted separately to each simulated subject's signed errors ($N \in \{1, 2, 4, 6, 8\}$, 500 trials per subject) by maximum likelihood with differential evolution (following van den Berg et al. [4]). Akaike information criterion (AIC), Bayesian information criterion (BIC), and the small-sample correction to AIC (AICc) were computed per subject (Supplementary Table S2).

In addition to comparing information criteria, we assessed whether each fitted model reproduced the mixture statistics of MemGP-generated errors. This comparison (Supplementary Fig. S2) followed van den Berg et al. [4]: for each subject and model, 10 synthetic datasets of the subject's size were generated from the maximum-likelihood parameters, the uniform–von Mises mixture was fit to each, and the estimated von Mises mixture weight w and component circular SD (CSD) were averaged to give that subject's prediction. Ribbons are ± s.e.m. of the predictions across subjects, and root mean squared error (RMSE) between predicted mixture parameters and those estimated from MemGP-generated errors was computed over all subjects and set sizes.

5.5 Statistical analysis

Error bars on behavioral simulation panels are s.e.m. across 20 simulated subjects that differ in random seed. Exceptions are the precision–repulsion analysis ($n = 24$, with encoding noise varied across subjects) and maintenance traces (Fig. 3k–l), which show means ± s.e.m. across 100 trials. Circuit trajectory bands are bootstrap 95% intervals as specified in Fig. 5.

Retrocue benefits (Fig. 3f) were paired t-tests on subject-mean absolute error (20 subjects, 100 trials per condition; cued vs. neutral). The load comparison was an independent-samples t-test on those subject-level benefits (one-tailed, H_1: greater benefit at $N = 4$ than at $N = 2$). The cue-reallocation test (Fig. 3h) was a paired t-test, across 20 subjects, of the trade-off $\Delta\mathrm{A} - \Delta\mathrm{B}$ (error change for the abandoned item minus that for the newly cued item), computed from each subject's mean errors over 100 trials. The maintenance-dynamics comparisons (Fig. 3k–l) were paired t-tests of the end-of-delay change from cue onset, cued versus the mean of the uncued items, across 100 trials: predictive mean in (k) and posterior variance in (l).

Bias used Eq. 6 with the simulated subject as the unit ($n = 20$, 100 trials per condition). Precision–repulsion correlations (Fig. 4f–g) are Pearson r on subject means. Other correlations

are Pearson r or Spearman ρ as labeled at each use, with 95% confidence intervals (Fisher z for Pearson; bootstrap for Spearman where noted). The uncertainty–error correlation in Fig. 5e uses 400 item-level observations from 100 displays; its reported interval resamples items and does not account for dependence between items from the same display.

Bootstrap standard errors ($B = 80$) were computed for mixture-model fits. The inferential tests reported above concern differences across simulated subjects and trials; they do not provide inferential evidence about human participants. Because simulation sample size depends on the number of trials generated, we emphasize effect sizes with uncertainty (95% CIs, or means ± s.e.m. for descriptive condition summaries).

Data and code availability

The code for the MemGP model, the simulations, the analyses, and the figures, together with the simulation outputs that the figures use, is available at `https://github.com/Daniel-Gong/MemGP-Code`. This study collected no new human data. The human data replotted in Figs. 2–4 come from published studies [3, 30, 45, 46]; copies of the files used, with their sources, are provided in the `data` folder of the same repository.

Author contributions

Conceptualization: D.G., M.B. and I.Y. Code: D.G. Formal analysis: D.G. Investigation: D.G. Visualization: D.G. Writing (original draft): D.G. and I.Y. Writing (review & editing): D.G., M.B. and I.Y. Supervision: I.Y.

Competing interests

The authors declare no competing interests.

References

[1] Steven J. Luck and Edward K. Vogel. The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657):279–281, November 1997. ISSN 1476-4687. doi: 10.1038/36846.

[2] Weiwei Zhang and Steven J. Luck. Discrete fixed-resolution representations in visual working memory. *Nature*, 453(7192):233–235, May 2008. ISSN 1476-4687. doi: 10.1038/nature06860.

[3] Paul M. Bays, Raquel F. G. Catalao, and Masud Husain. The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, 9(10):7, September 2009. ISSN 1534-7362. doi: 10.1167/9.10.7.

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

[5] Wei Ji Ma, Masud Husain, and Paul M. Bays. Changing concepts of working memory. *Nature Neuroscience*, 17(3):347–356, March 2014. ISSN 1546-1726. doi: 10.1038/nn.3655.

[6] Xiao-Jing Wang. Synaptic reverberation underlying mnemonic persistent activity. *Trends in neurosciences*, 24(8):455–463, 2001.

[7] Paul M. Bays. Noise in Neural Populations Accounts for Errors in Working Memory. *Journal of Neuroscience*, 34(10):3632–3645, March 2014. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.3204-13.2014.

[8] Flora Bouchacourt and Timothy J Buschman. A flexible model of working memory. *Neuron*, 103(1):147–160, 2019.

[9] Gianluigi Mongillo, Omri Barak, and Misha Tsodyks. Synaptic theory of working memory. *Science*, 319(5869):1543–1546, 2008. doi: 10.1126/science.1150769.

[10] Mark G. Stokes. 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences*, 19(7):394–405, July 2015. ISSN 1364-6613. doi: 10.1016/j.tics.2015.05.004.

[11] Bradley R. Postle and Klaus Oberauer. Working Memory: Theoretical, Computational, and Neural Considerations. In Michael J. Kahana and Anthony D. Wagner, editors, *The Oxford Handbook of Human Memory, Two Volume Pack*, pages 371–414. Oxford University Press, 1 edition, July 2024. ISBN 978-0-19-091798-2 978-0-19-091801-9. doi: 10.1093/oxfordhb/9780190917982.013.14.

[12] Paul M. Bays, Sebastian Schneegans, Wei Ji Ma, and Timothy F. Brady. Representation and computation in visual working memory. *Nature Human Behaviour*, 8(6):1016–1034, June 2024. ISSN 2397-3374. doi: 10.1038/s41562-024-01871-2.

[13] Ziqiang Wei, Xiao-Jing Wang, and Da-Hui Wang. From distributed resources to limited slots in multiple-item working memory: A spiking network model with normalization. *Journal of Neuroscience*, 32(33):11228–11240, 2012. doi: 10.1523/JNEUROSCI.0735-12.2012.

[14] Jeffrey S. Johnson, Vanessa R. Simmering, and Aaron T. Buss. Beyond slots and resources: Grounding cognitive concepts in neural dynamics. *Attention, Perception, & Psychophysics*, 76 (6):1630–1654, 2014. doi: 10.3758/s13414-013-0596-9.

[15] Loic Matthey, Paul M. Bays, and Peter Dayan. A probabilistic palimpsest model of visual short-term memory. *PLoS Computational Biology*, 11(1):e1004003, 2015. doi: 10.1371/journal.pcbi.1004003.

[16] Paul M. Bays. Spikes not slots: Noise in neural populations limits working memory. *Trends in Cognitive Sciences*, 19(8):431–438, August 2015. ISSN 1879-307X. doi: 10.1016/j.tics.2015.06.004.

[17] Sebastian Schneegans, Robert Taylor, and Paul M. Bays. Stochastic sampling provides a unifying account of visual working memory limits. *Proceedings of the National Academy of Sciences*, 117(34):20959–20968, August 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2004306117.

[18] Mark W. Schurgin, John T. Wixted, and Timothy F. Brady. Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour*, 4(11):1156–1172, November 2020. ISSN 2397-3374. doi: 10.1038/s41562-020-00938-0.

[19] David C. Knill and Alexandre Pouget. The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712–719, December 2004. ISSN 01662236. doi: 10.1016/j.tins.2004.10.007.

[20] Wei Ji Ma and Mehrdad Jazayeri. Neural Coding of Uncertainty and Probability. *Annual Review of Neuroscience*, 37:205–220, 2014. ISSN 1545-4126. doi: 10.1146/annurev-neuro-071013-014017.

[21] Hsin-Hung Li, Thomas C. Sprague, Aspen H. Yoo, Wei Ji Ma, and Clayton E. Curtis. Joint representation of working memory and uncertainty in human cortex. *Neuron*, 109(22):3699–3712.e6, November 2021. ISSN 0896-6273. doi: 10.1016/j.neuron.2021.08.022.

[22] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press, Cambridge, MA, 2006.

[23] Michalis K. Titsias. Variational learning of inducing variables in sparse gaussian processes. In David van Dyk and Max Welling, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574, Clearwater Beach, Florida, USA, 2009. PMLR. URL https://proceedings.mlr.press/v5/titsias09a.html.

[24] Felix Leibfried, Vincent Dutordoir, S. T. John, and Nicolas Durrande. A Tutorial on Sparse Gaussian Processes and Variational Inference, December 2022.

[25] Ronald van den Berg and Wei Ji Ma. A resource-rational theory of set size effects in human visual working memory. *eLife*, 7:e34963, August 2018. ISSN 2050-084X. doi: 10.7554/eLife.34963.

[26] Matteo Carandini and David J. Heeger. Summation and Division by Neurons in Primate Visual Cortex. *Science*, 264(5163):1333–1336, May 1994. doi: 10.1126/science.8191289.

[27] Matteo Carandini and David J. Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2012. doi: 10.1038/nrn3136.

[28] Klaus Oberauer and Hsuan-Yu Lin. An interference model of visual working memory. *Psychological Review*, 124(1):21–59, 2017. ISSN 1939-1471. doi: 10.1037/rev0000044.

[29] Gi-Yeul Bae and Steven J. Luck. Interactions between visual working memory representations. *Attention, Perception, & Psychophysics*, 79(8):2376–2395, August 2017. ISSN 1943-393X. doi: 10.3758/s13414-017-1404-8.

[30] Chaipat Chunharas, Rosanne L. Rademaker, Timothy F. Brady, and John T. Serences. An adaptive perspective on visual working memory distortions. *Journal of Experimental Psychology: General*, 151(10):2300–2323, 2022. ISSN 1939-2222. doi: 10.1037/xge0001191.

[31] Patrick Wilken and Wei Ji Ma. A detection theory account of change detection. *Journal of Vision*, 4(12):1120–1135, December 2004. ISSN 1534-7362. doi: 10.1167/4.12.11.

[32] Ivan C. Griffin and Anna C. Nobre. Orienting Attention to Locations in Internal Representations. *Journal of Cognitive Neuroscience*, 15(8):1176–1194, November 2003. doi: 10.1162/089892903322598139.

[33] Alessandra S. Souza and Klaus Oberauer. In search of the focus of attention in working memory: 13 years of the retro-cue effect. *Attention, Perception, & Psychophysics*, 78(7):1839–1860, 2016. doi: 10.3758/s13414-016-1108-5.

[34] Zachary Lively, Maria M. Robinson, and Aaron S. Benjamin. Memory fidelity reveals qualitative changes in interactions between items in visual working memory. *Psychological Science*, 32(9): 1426–1441, 2021. doi: 10.1177/0956797621997367.

[35] Albert Compte, Nicolas Brunel, Patricia S. Goldman-Rakic, and Xiao-Jing Wang. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, 10(9):910–923, 2000. doi: 10.1093/cercor/10.9.910.

[36] Klaus Wimmer, Duane Q. Nykamp, Christos Constantinidis, and Albert Compte. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature Neuroscience*, 17(3):431–439, March 2014. ISSN 1546-1726. doi: 10.1038/nn.3645.

[37] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

[38] Stephen M. Emrich and Susanne Ferber. Competition increases binding errors in visual working memory. *Journal of Vision*, 12(4):12, April 2012. ISSN 1534-7362. doi: 10.1167/12.4.12.

[39] Laura Rerko, Klaus Oberauer, and Hsuan-Yu Lin. Spatial transposition gradients in visual working memory. *Quarterly Journal of Experimental Psychology*, 67(1):3–15, January 2014. ISSN 1747-0226. doi: 10.1080/17470218.2013.789543.

[40] Paul M. Bays. Evaluating and excluding swap errors in analogue tests of working memory. *Scientific Reports*, 6(1):19203, January 2016. ISSN 2045-2322. doi: 10.1038/srep19203.

[41] Sebastian Schneegans and Paul M. Bays. Neural Architecture for Feature Binding in Visual Working Memory. *The Journal of Neuroscience*, 37(14):3913–3925, April 2017. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.3493-16.2017.

[42] Jessica M. V. McMaster, Ivan Tomić, Sebastian Schneegans, and Paul M. Bays. Swap errors in visual working memory are fully explained by cue-feature variability. *Cognitive Psychology*, 137:101493, September 2022. ISSN 0010-0285. doi: 10.1016/j.cogpsych.2022.101493.

[43] Daryl Fougnie, Jordan W. Suchow, and George A. Alvarez. Variability in the quality of visual working memory. *Nature Communications*, 3(1):1229, November 2012. ISSN 2041-1723. doi: 10.1038/ncomms2237.

[44] Freek Van Ede and Anna C Nobre. Turning attention inside out: How working memory serves behavior. *Annual review of psychology*, 74(1):137–165, 2023.

[45] Alessandra S. Souza, Laura Rerko, Hsuan-Yu Lin, and Klaus Oberauer. Focused attention improves working memory: Implications for flexible-resource and discrete-capacity models. *Attention, Perception, & Psychophysics*, 76(7):2085–2104, 2014. doi: 10.3758/s13414-014-0687-2.

[46] Dirk van Moorselaar, Christian N. L. Olivers, Jan Theeuwes, Victor A. F. Lamme, and Ilja G. Sligte. Forgotten but not gone: Retro-cue costs and benefits in a double-cueing paradigm suggest multiple states in visual short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(6):1755–1763, 2015. doi: 10.1037/xlm0000124.

[47] Yiqing Hu, Xuye Yuan, Yiwen Li, Yuanjun Kong, Jing Huang, Yan Song, and Dongwei Li. Load-Dependent Retrospective Attentional Benefits During Visual Working Memory. *Neuroscience Bulletin*, December 2025. ISSN 1995-8218. doi: 10.1007/s12264-025-01556-4.

[48] Timothy F. Brady and George A. Alvarez. Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, 22(3):384–392, 2011. doi: 10.1177/0956797610397956.

[49] Paul S. Scotti, Yoolim Hong, Andrew B. Leber, and Julie D. Golomb. Visual working memory items drift apart due to active, not passive, maintenance. *Journal of Experimental Psychology: General*, 150(12):2506–2524, 2021. doi: 10.1037/xge0000890.

[50] Igor S. Utochkin and Timothy F. Brady. Individual representations in visual working memory inherit ensemble properties. *Journal of Experimental Psychology: Human Perception and Performance*, 46(5):458–473, 2020. doi: 10.1037/xhp0000727.

[51] Anxin Miao, Timothy F. Brady, and Maria M. Robinson. Bridging Bayesian and representational theories of memory to predict memory bias. *Journal of Experimental Psychology: General*, 155 (11):2631–2644, 2026. ISSN 1939-2222. doi: 10.1037/xge0001945.

[52] Garrett Swan and Brad Wyble. The binding pool: A model of shared neural resources for distinct items in visual working memory. *Attention, Perception, & Psychophysics*, 76(7): 2136–2157, October 2014. ISSN 1943-393X. doi: 10.3758/s13414-014-0633-3.

[53] Rita Almeida, João Barbosa, and Albert Compte. Neural circuit basis of visuo-spatial working memory precision: a computational and behavioral study. *Journal of neurophysiology*, 114(3): 1806–1818, 2015.

[54] Geoffrey E. Hinton and David C. Plaut. Using fast weights to deblur old memories. In *Proceedings of the 9th Annual Conference of the Cognitive Science Society*, pages 177–186, Hillsdale, NJ, 1987. Lawrence Erlbaum Associates.

[55] Jürgen Schmidhuber. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139, 1992.

[56] Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International conference on machine learning*, pages 9355–9366. PMLR, 2021.

[57] Ruben S. van Bergen, Wei Ji Ma, Michael S. Pratte, and Janneke F. M. Jehee. Sensory uncertainty decoded from visual cortex predicts behavior. *Nature Neuroscience*, 18(12):1728–1730, December 2015. ISSN 1546-1726. doi: 10.1038/nn.4150.

[58] Thomas Parr and Karl J. Friston. Working memory, attention, and salience in active inference. *Scientific Reports*, 7(1):14678, November 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-15249-0.

[59] Samuel J Gershman and Jeffrey M Beck. Complex probabilistic inference: From cognition to neural computation. *Computational models of brain and behavior*, pages 453–466, 2017.

[60] Hadi Vafaii, Dekel Galor, and Jacob Yates. Brain-like variational inference. *Advances in Neural Information Processing Systems*, 38:163700–163778, 2026.

[61] Hadi Vafaii and Jacob L. Yates. Metabolic cost of information processing in Poisson variational autoencoders. In *Proceedings of Cognitive Computational Neuroscience 2026*, New York City, 2026. Cognitive Computational Neuroscience. doi: 10.32470/6ff31r0.

[62] Anna Kutschireiter, Melanie A Basnak, Rachel I Wilson, and Jan Drugowitsch. Bayesian inference in ring attractor networks. *Proceedings of the National Academy of Sciences*, 120(9): e2210622120, 2023.

[63] Shun-ichi Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27(2):77–87, 1977. doi: 10.1007/BF00337259.

[64] Shintaro Funahashi, Charles J. Bruce, and Patricia S. Goldman-Rakic. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, 61(2): 331–349, February 1989. ISSN 0022-3077. doi: 10.1152/jn.1989.61.2.331.

[65] S. Chenchal Rao, Gregor Rainer, and Earl K. Miller. Integration of what and where in the primate prefrontal cortex. *Science*, 276(5313):821–824, May 1997. ISSN 1095-9203. doi: 10.1126/science.276.5313.821.

[66] Omri Barak and Misha Tsodyks. Persistent activity in neural networks with dynamic synapses. *PLoS Computational Biology*, 3(2):e35, 2007. doi: 10.1371/journal.pcbi.0030035.

[67] Michael J. Wolff, Janina Jochim, Elkan G. Akyürek, and Mark G. Stokes. Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience*, 20(6):864–871, 2017. doi: 10.1038/nn.4546.

[68] Nicolas Y. Masse, Guangyu R. Yang, H. Francis Song, Xiao-Jing Wang, and David J. Freedman. Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nature Neuroscience*, 22(7):1159–1167, 2019. doi: 10.1038/s41593-019-0414-3.

[69] Joao Barbosa, Heike Stein, Rebecca L. Martinez, Adria Galan-Gadea, Sihai Li, Josep Dalmau, Kirsten C. S. Adam, Josep Valls-Solé, Christos Constantinidis, and Albert Compte. Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. *Nature Neuroscience*, 23(8):1016–1024, 2020. doi: 10.1038/s41593-020-0644-4.

[70] Jochen Görtler, Rebecca Kehlbeck, and Oliver Deussen. A Visual Exploration of Gaussian Processes. *Distill*, 4(4):e17, April 2019. ISSN 2476-0757. doi: 10.23915/distill.00017.

[71] Matthew F. Panichello, Brian DePasquale, Jonathan W. Pillow, and Timothy J. Buschman. Error-correcting dynamics in visual working memory. *Nature Communications*, 10:3366, 2019. doi: 10.1038/s41467-019-11298-3.

[72] Wei Ji Ma, Jeffrey M. Beck, Peter E. Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438, 2006. doi: 10.1038/nn1790.

[73] Charley M. Wu, Eric Schulz, Maarten Speekenbrink, Jonathan D. Nelson, and Björn Meder. Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, 2 (12):915–924, November 2018. ISSN 2397-3374. doi: 10.1038/s41562-018-0467-4.

[74] Eric Schulz, Maarten Speekenbrink, and Andreas Krause. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, 85:1–16, August 2018. ISSN 0022-2496. doi: 10.1016/j.jmp.2018.03.001.

[75] Jasper Snoek, Hugo Larochelle, and Ryan Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

[76] Jacob R. Gardner, Geoff Pleiss, Kilian Q. Weinberger, David Bindel, and Andrew Gordon Wilson. GPyTorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Supplementary Materials

S1. Mathematical formulation

S1.1 Gaussian process prior

The latent memory-strength function $f : \mathcal{X} \to \mathbb{R}$ assigns a strength to each location–color combination, where $\mathcal{X} = [-180, 180)^2$ is the circular location–color space and each input is a pair $\mathbf{x} = (\ell, c)$. The GP prior (Eq. 1) is

$$f \sim \mathcal{GP}\big(m(\mathbf{x}),\, k(\mathbf{x}, \mathbf{x}')\big)$$

where $m(\mathbf{x})$ is the prior mean, constant across feature space and learned during inference. The prior covariance between two location–color combinations is

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \cdot k_{\text{per}}(\ell, \ell'; \lambda_\ell, P) \cdot k_{\text{per}}(c, c'; \lambda_c, P) \tag{S1}$$

The covariance is the product of a location kernel and a color kernel, multiplied by the output scale σ_f^2. Each periodic kernel is

$$k_{\text{per}}(x, x'; \lambda, P) = \exp\left(-\frac{2\sin^2(\pi|x - x'|/P)}{\lambda}\right) \tag{S2}$$

with period $P = 360°$ for both dimensions. The denominator follows the GPyTorch PeriodicKernel convention: it contains λ, not its square. The parameter λ determines how covariance changes with distance along the circular dimension. The output scale σ_f^2 is learned as part of a `ScaleKernel` wrapper. Figure S1 shows independent draws of f from $\mathcal{GP}(0, k)$.

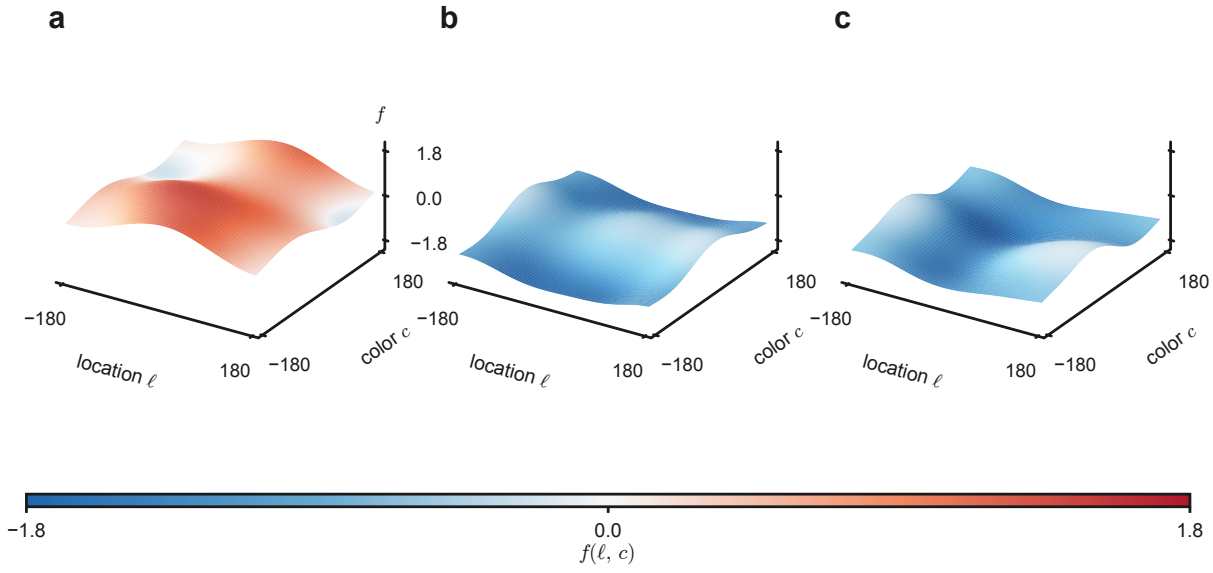


Figure S1: The GP prior is a smooth random function on the circular location–color torus. **(a–c)** Independent zero-mean draws $f \sim \mathcal{GP}(0, k)$, each a map $[-180, 180)^2 \to \mathbb{R}$. Height and color encode f.

S1.2 Sparse variational approximation

An observation consists of a location–color combination $\mathbf{x}_i$ and an evidence strength y_i. The Gaussian likelihood $p(y_i \mid f_i)$ relates the observed strength to the latent function value $f_i = f(\mathbf{x}_i)$. The exact posterior at a query location $\mathbf{x}^*$ is $p(f^* \mid \mathbf{X}, \mathbf{y}, \mathbf{x}^*)$, $f^* = f(\mathbf{x}^*)$. We approximate it using M inducing points $\mathbf{Z} = \{(\ell_m^z, c_m^z)\}_{m=1}^M$, $\mathbf{Z} \in \mathbb{R}^{M\times 2}$. The inducing values are $\mathbf{u} = f(\mathbf{Z})$. The variational mean $\mathbf{m} \in \mathbb{R}^M$ and covariance $\mathbf{S} \in \mathbb{R}^{M\times M}$ describe the distribution of these inducing values. The diagonal entries of $\mathbf{S}$ are marginal variances; off-diagonal entries are covariances between inducing values. The implementation parameterizes $\mathbf{S}$ through a Cholesky factor, whose product with its transpose gives the covariance (`CholeskyVariationalDistribution`). The variational distribution (Eq. 2) is

$$q(\mathbf{u}) = \mathcal{N}(\mathbf{u};\, \mathbf{m},\, \mathbf{S}).$$

To predict strength at a queried location–color combination, the model integrates the GP conditional over the distribution of inducing values. This gives the approximate posterior predictive distribution:

$$q(f(\mathbf{x}^*)) = \mathcal{N}\Big(\mu_0 + \mathbf{k}_*^\top \mathbf{K}_{ZZ}^{-1}(\mathbf{m} - \mu_0 \mathbf{1}),\; k_{**} - \mathbf{k}_*^\top \mathbf{K}_{ZZ}^{-1}(\mathbf{K}_{ZZ} - \mathbf{S})\mathbf{K}_{ZZ}^{-1}\mathbf{k}_*\Big) \tag{S3}$$

Here μ_0 is the constant prior mean. The vector $\mathbf{k}_* = k(\mathbf{Z}, \mathbf{x}^*)$ contains prior covariances between the queried combination and the inducing points; $k_{**} = k(\mathbf{x}^*, \mathbf{x}^*)$ is the prior variance at the query. The matrix $\mathbf{K}_{ZZ} = k(\mathbf{Z}, \mathbf{Z})$ contains prior covariances among the inducing values. The first argument of the Gaussian is the predicted mean strength, and the second is its variance.

Inducing points are initialized on a regular $m \times m$ grid ($m = 10$ by default, $M = 100$) with centers offset by half the bin width to ensure uniform coverage. Their positions are inferred during optimization (`learn_inducing_locations = True`).

S1.3 Encoding objective

Given N items $\{(\check{\ell}_i, \check{c}_i)\}_{i=1}^N$, the surrogate perceptual procedure generates n_{obs} observations per item $\mathbf{x}_{ij} = (\check{\ell}_i + \epsilon_\ell, \check{c}_i + \epsilon_c)$ with $\epsilon_\ell \sim \mathcal{N}(0, \sigma_\ell^2)$, $\epsilon_c \sim \mathcal{N}(0, \sigma_c^2)$, wrapped to $[-180, 180)$. The evidence-strength targets $y_{ij} = w_{ij}$ are

$$w_{ij} = \exp\left(-\frac{d(\ell_{ij}, \check{\ell}_i)^2}{2\sigma_\ell^2}\right) \cdot \exp\left(-\frac{d(c_{ij}, \check{c}_i)^2}{2\sigma_c^2}\right), \tag{S4}$$

where $d(\cdot,\cdot)$ is circular distance.

The encoding ELBO is maximized using Adam with learning rate $\eta_{\text{enc}} = 0.05$ for $T_{\text{enc}} = 100$ epochs. The Gaussian likelihood noise variance is fixed (not inferred; Table S1).

S1.4 Maintenance objective

During maintenance, the model uses its preceding predictions as observations in the variational objective. The likelihood is evaluated at the current inducing locations $\mathbf{Z}^{(t)} = \{\mathbf{z}_m^{(t)}\}_{m=1}^M$. Self-

rehearsal targets are initialized at the end of encoding as the posterior mean of inducing values and then updated after each epoch from the current posterior mean,

$$\tilde{y}_m^{(0)} = \mathbb{E}_{q(f)}[f(\mathbf{z}_m)]\big|_{t=T_{\text{enc}}}, \qquad \tilde{y}_m^{(t)} = \mathbb{E}_{q(f)}[f(\mathbf{z}_m^{(t)})]\big|_{t-1}. \tag{S5}$$

Attention gains $a(\ell_m^z)$ are recomputed on the current inducing locations after cue onset (Eq. 7; Fig. 3).

The maintenance objective (Eq. 5) is optimized with Adam (learning rates and epoch counts in Table S1). The KL coefficient β controls regularization toward the prior and is set to $\beta = 2$ in the baseline configurations. The effect of KL regularization on posterior variance depends on the rehearsal term, the attention weights and the learned prior; it does not require variance to increase at every point on every update. Before cue onset (epoch t_{cue}), attention weights are uniform ($a(\ell_m^z) = 1$). After cue onset, the spatial gain function (Eq. 7) with parameters G (attended gain) and σ_s (spatial width) is applied.

S1.5 Retrieval

A single joint sample $f \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is drawn from the GP posterior predictive over 360 evenly spaced color values at every occupied item location, preserving inter-color and inter-item covariance. The sampled memory-strength profile over color at the probed location, $f_t(c) = f(\ell_{\text{probe}}, c)$, is then divisively normalized by a pool aggregated across all occupied item locations. Each item i at location $\check{\ell}_i$ contributes its sampled memory-strength profile over color $f_i(c) = f(\check{\ell}_i, c)$ from the same joint sample, weighted by a spatial surround weight,

$$w_i = \exp\!\big(-\, d(\check{\ell}_i, \ell_{\text{probe}})^2 \,/\, 2\sigma_{\text{pool}}^2\big)\,, \tag{S6}$$

The probed item has zero spatial distance from the probe and therefore enters with $w = 1$. Subtracting the same minimum sampled strength from every value, $f_i^+(c) = f_i(c) - \min f$, makes all responses nonnegative. The minimum is taken over every sampled item location and color; subtracting it rather than rectifying ensures that every item contributes a nonnegative response to the pool, including items with little evidence. The pool is the weighted sum of shifted responses:

$$\text{pool}(c) = \textstyle\sum_i w_i\, f_i^+(c), \tag{S7}$$

For each candidate color, the probed item's shifted strength is divided by this pool plus a baseline term. The normalized response and the selected report are

$$r(c) = \frac{f_t^+(c)}{\sigma_b \cdot \max_c f_t^+(c) + \text{pool}(c)}, \qquad \hat{c} = \arg\max_c\, r(c), \tag{S8}$$

where σ_b is the baseline (semi-saturation) fraction. Two parameters govern the readout: the normalization-pool spatial width σ_{pool}, which determines how rapidly each item's weight w_i decreases

with distance from the probe, and the baseline fraction σ_b, which scales the maximum target response in the denominator (Table S1). The baseline term scales with the maximum shifted response of the probed item rather than being a fixed constant as in the standard normalization equation [27]; this makes the readout invariant to the overall scale of the sampled strengths, which decline during the delay.

S1.6 Misbinding as a function of set size and target–distractor distance

To test whether swap errors are spatially structured, we simulated set sizes $N \in \{2, 4, 6, 8\}$ and probed every item as the target, drawing 100 posterior-sample recalls per probe. A recall was counted as a swap toward a given non-target item if it fell within 30° of that item's color but not within 30° of the target's color. For every target–distractor pair we computed the circular spatial distance $d(\ell_{\text{probe}}, \check{\ell}_j)$, which takes the discrete values $\{45°, 90°, 135°, 180°\}$ under the eight-position stimulus geometry, and grouped pairs by both set size and distance. The heatmap reports the mean misbinding rate per (set size, distance) cell. Misbinding decreases with distance and increases with set size.

Table S1: Baseline settings across behavioral experiments.

Parameter	Set-size	Retrocue	Bias
n_{obs} per item	1000	1000	1000
σ_ℓ (loc sampling noise)	20°	20°	20°
σ_c (color sampling noise)	20°	20°	20°
m (inducing grid side)	10	10	10
λ_ℓ (initial loc lengthscale)	10.0	10.0	10.0
λ_c (initial color lengthscale)	10.0	10.0	10.0
σ_n^2 (likelihood noise)	10^{-4}	10^{-4}	10^{-4}
T_{enc} (encoding epochs)	100	100	100
T_{maint} (maintenance epochs)	100	100[a]	100
η_{enc} (encoding LR)	0.05	0.05	0.05
η_{maint} (maintenance LR)	0.01	0.01	0.01
β (KL weight)	2	2	2
σ_s (attention spatial width)	–	20°	–
G (attended gain)	–	40.0	–
σ_{pool} (normalization-pool spatial width)	60°	60°	60°
σ_b (baseline fraction)	0.8	0.8	0.8

Inducing positions, kernel lengthscales, and prior mean are learned during inference. Bias trials with a target cue use the retrocue attention settings. [a]The cue-flexibility simulation (Fig. 3h) uses $T_{\text{maint}} = 200$ to accommodate the two cue windows (cue A, epochs 50–99; cue B, epochs 100–199).

S2. Information-theoretic model comparison

Following van den Berg et al. [4], we compared descriptive models fitted to MemGP errors. For each descriptive model $\mathcal{M}$, we estimated k free parameters by maximum likelihood estimation (MLE) from n recall errors. We compared fits using Akaike information criterion (AIC), Bayesian

information criterion (BIC), and the small-sample correction to AIC (AICc):

$$\mathrm{AIC} = 2k - 2\ln\hat{L} \tag{S9}$$

$$\mathrm{BIC} = k\ln n - 2\ln\hat{L} \tag{S10}$$

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1} \tag{S11}$$

Here $\hat{L}$ is the maximized likelihood. Each criterion combines a term measuring fit with a penalty for the number of fitted parameters. Lower AIC/BIC indicates a preferred model under the corresponding criterion (see Table S2). Parameter counts: IL ($k = 3$: $K, \kappa_{\mathrm{mem}}, \kappa_r$), SA ($k = 3$: K, J_1, κ_r), EP ($k = 3$: J_1, α, κ_r), VP ($k = 4$: $\bar{J}_1, \alpha, \tau, \kappa_r$).

Table S2: Information-theoretic model comparison. Mean ΔAIC and ΔBIC (± s.e.m.) across 20 simulated subjects. Each model was fit by MLE to that subject's signed errors ($n = 500$: 100 trials × set sizes 1, 2, 4, 6, 8). Differences are from the best-fitting model in that subject. VP had the lowest AIC and BIC in all 20 simulated subjects. These criteria compare descriptive models fitted to MemGP-generated recall errors.

Model	k	**ΔAIC**	**ΔBIC**	**Best (AIC)**
VP	4	0.0 ± 0.0	0.0 ± 0.0	20/20
IL	3	89.0 ± 3.5	84.8 ± 3.5	0/20
SA	3	91.7 ± 7.2	87.5 ± 7.2	0/20
EP	3	131.8 ± 8.7	127.6 ± 8.7	0/20

S3. Recurrent circuit implementation of MemGP

To derive recurrent equations, we hold the inducing locations, kernel parameters, likelihood noise and prior mean fixed. Only the variational mean and covariance are updated in the derivation. We express the variational objective in terms of the unwhitened mean $\mathbf{m}$ and covariance $\mathbf{S}$ of the inducing values $\mathbf{u} = f(\mathbf{Z})$. Unwhitened coordinates describe the original function values; the numerical implementation uses transformed coordinates, defined in S3.3. The inverse prior covariance, or prior precision, determines interactions between units that remain fixed under these assumptions, while $\mathbf{S}$ is a rapidly updated covariance state. Attention changes diagonal gains in the effective operator $\mathbf{W}$. Interpreting these states as rates and fast weights specifies a candidate computational implementation; the equations do not establish a local synaptic learning rule.

The likelihood is Gaussian, $p(y \mid f) = \mathcal{N}(y; f, \sigma^2)$, and the prior on the inducing values is $p(\mathbf{u}) = \mathcal{N}(\mu_0\mathbf{1}, \mathbf{K}_{ZZ})$, with μ_0 the constant prior mean. The variational posterior is $q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, \mathbf{S})$. For one latent strength with marginal posterior $q(f_i) = \mathcal{N}(\mu_i, v_i)$, μ_i is the posterior mean and v_i is the posterior variance. Averaging the log-likelihood over uncertainty in that strength gives

$$\mathbb{E}_q[\log p(y_i \mid f_i)] = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - \mu_i)^2 + v_i}{2\sigma^2}. \tag{S12}$$

The KL divergence measures departure of the Gaussian posterior $q(\mathbf{u})$ from the Gaussian prior

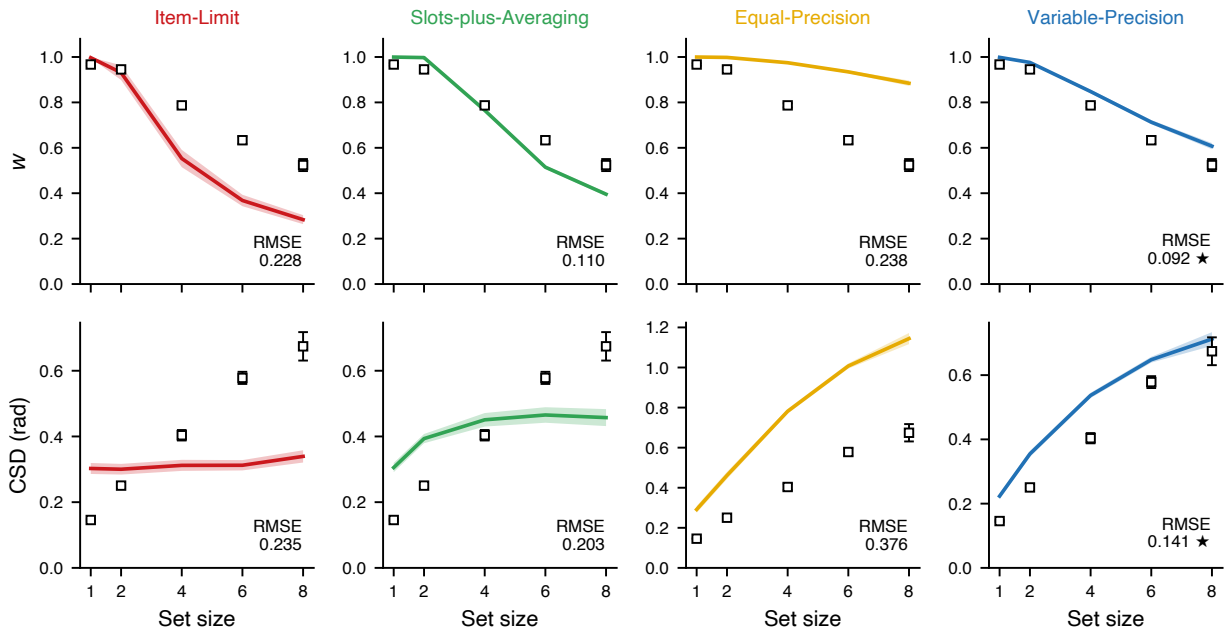


Figure S2: Formal model comparison. Uniform–von Mises mixture statistics extracted from MemGP simulation errors (open squares, mean ± s.e.m. across 20 simulated subjects) compared with predictions from four benchmark models fit by maximum likelihood to each subject's errors ($N \in \{1, 2, 4, 6, 8\}$, 100 trials each). Top row: mixture weight w (probability of non-guess response). Bottom row: circular standard deviation (CSD) of the von Mises component. Colored lines show mean model predictions across subjects; shaded ribbons are ± s.e.m. across subjects. RMSE, computed over all subjects and set sizes between each subject's prediction and statistics, quantifies goodness of fit; ★ marks the lowest RMSE. VP had the lowest RMSE for both mixture statistics and the lowest AIC and BIC (Supplementary Table S2).

$p(\mathbf{u})$. Its closed-form expression is

$$\begin{aligned}\mathrm{KL}\big(q(\mathbf{u}) \,\|\, p(\mathbf{u})\big) = \frac{1}{2}(\mathbf{m}-\mu_0\mathbf{1})^\top \mathbf{K}_{ZZ}^{-1}(\mathbf{m}-\mu_0\mathbf{1}) + \frac{1}{2}\mathrm{tr}(\mathbf{K}_{ZZ}^{-1}\mathbf{S}) \\ - \frac{1}{2}\log|\mathbf{S}| + \frac{1}{2}\log|\mathbf{K}_{ZZ}| - \frac{M}{2}.\end{aligned} \tag{S13}$$

S3.1 Encoding

For n perceptual observations, Eq. S12 gives the negative expected log-likelihood in Eq. 4:

$$-\sum_{i=1}^{n}\mathbb{E}_q[\log p(y_i \mid f_i)] = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\|\mathbf{y}-\boldsymbol{\mu}\|^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n} v_i. \tag{S14}$$

To differentiate the encoding objective with respect to the inducing mean, we first express the predictions at observation locations in terms of that mean. Let $\boldsymbol{\Phi} = \mathbf{K}_{XZ}\mathbf{K}_{ZZ}^{-1}$, where $\mathbf{K}_{XZ}$ contains prior covariances between observation locations and inducing points. Multiplying by $\boldsymbol{\Phi}$ maps inducing-value deviations from the prior mean to predicted deviations at the observation locations. For a constant prior mean, the predictive mean in Eq. S3 is $\boldsymbol{\mu} = \mu_0\mathbf{1}_n + \boldsymbol{\Phi}(\mathbf{m}-\mu_0\mathbf{1}_M)$. Define the adjusted observation vector $\mathbf{y}_c = \mathbf{y} - \mu_0\mathbf{1}_n + \boldsymbol{\Phi}\mu_0\mathbf{1}_M$, so that $\mathbf{y}-\boldsymbol{\mu} = \mathbf{y}_c - \boldsymbol{\Phi}\mathbf{m}$. The predictive variance at $\mathbf{x}_i$ is

$$v_i = k(\mathbf{x}_i,\mathbf{x}_i) - \mathbf{k}_i^\top \mathbf{K}_{ZZ}^{-1}\mathbf{k}_i + \mathbf{k}_i^\top \mathbf{K}_{ZZ}^{-1}\mathbf{S}\,\mathbf{K}_{ZZ}^{-1}\mathbf{k}_i, \tag{S15}$$

Here $\mathbf{k}_i$ is the vector of prior covariances between observation $\mathbf{x}_i$ and the inducing points. The predictive variance depends on $\mathbf{S}$ and on the kernel, but is independent of $\mathbf{m}$.

Define $\bar{\mathcal{L}}_{\mathrm{enc}} = \frac{1}{n}\mathcal{L}_{\mathrm{enc}}$, substitute (S14) and (S13) into (4), and keep the terms that contain $\mathbf{m}$:

$$\begin{aligned}\bar{\mathcal{L}}_{\mathrm{enc}}(\mathbf{m}) &= \frac{1}{n}\left(\frac{1}{2\sigma^2}\|\mathbf{y}_c-\boldsymbol{\Phi}\mathbf{m}\|^2 + \frac{1}{2}(\mathbf{m}-\mu_0\mathbf{1})^\top\mathbf{K}_{ZZ}^{-1}(\mathbf{m}-\mu_0\mathbf{1})\right) + \text{const} \\ &= \frac{1}{2\sigma^2 n}\|\mathbf{y}_c-\boldsymbol{\Phi}\mathbf{m}\|^2 + \frac{1}{2n}(\mathbf{m}-\mu_0\mathbf{1})^\top\mathbf{K}_{ZZ}^{-1}(\mathbf{m}-\mu_0\mathbf{1}) + \text{const}.\end{aligned} \tag{S16}$$

where the constant collects everything independent of $\mathbf{m}$, including $\sum_i v_i$, $\mathrm{tr}(\mathbf{K}_{ZZ}^{-1}\mathbf{S})$, and $\log|\mathbf{S}|$.

Differentiating (S16) with respect to $\mathbf{m}$,

$$\nabla_{\mathbf{m}}\bar{\mathcal{L}}_{\mathrm{enc}} = \frac{1}{\sigma^2 n}\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\mathbf{m}-\mathbf{y}_c) + \frac{1}{n}\mathbf{K}_{ZZ}^{-1}(\mathbf{m}-\mu_0\mathbf{1}) = \mathbf{W}_{\mathrm{enc}}\mathbf{m} - \mathbf{b}_{\mathrm{enc}}, \tag{S17}$$

where

$$\mathbf{W}_{\mathrm{enc}} = \frac{1}{\sigma^2 n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \frac{1}{n}\mathbf{K}_{ZZ}^{-1}, \qquad \mathbf{b}_{\mathrm{enc}} = \frac{1}{\sigma^2 n}\boldsymbol{\Phi}^\top\mathbf{y}_c + \frac{1}{n}\mathbf{K}_{ZZ}^{-1}\mu_0\mathbf{1}. \tag{S18}$$

Gradient flow on Eq. S16 gives $\tau\dot{\mathbf{m}} = -\mathbf{W}_{\mathrm{enc}}\mathbf{m} + \mathbf{b}_{\mathrm{enc}}$ in the recurrent circuit. The input drive includes $\boldsymbol{\Phi}^\top\mathbf{y}_c$, which projects the adjusted observations onto the inducing units. The recurrent operator contains a likelihood term determined by $\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ and a prior term determined by the

precision matrix. Because $\mathbf{\Phi}^\top \mathbf{\Phi}$ is positive semidefinite and $\mathbf{K}_{ZZ}^{-1}$ is positive definite, $\mathbf{W}_{\text{enc}} \succ 0$, and $\mathbf{m}$ converges to the unique minimizer of (S16) from every initial value. This convergence statement assumes a fixed objective and applies to the ideal gradient flow; finite Adam updates need not reach its equilibrium.

S3.2 Maintenance

During maintenance, the encoding observations are replaced by the network's preceding predictions. The maintenance objective (Eq. 5) therefore weights the expected log-likelihood of the network's own previous posterior means by the attention gains, and multiplies the KL by β. The rehearsal targets $\mathbf{t} = \tilde{\mathbf{y}}$ are the posterior mean strengths from the preceding epoch. Each epoch holds these targets fixed while updating the variational mean and covariance, then replaces the targets with the updated predicted means. The maintenance likelihood is evaluated at the inducing points, where the unwhitened predictive moments are $\mathbf{m}$ and $\mathbf{S}$. Let $\mathbf{A} = \operatorname{diag} a(\ell_m^{(z)})$ be the diagonal matrix of attention gains, with one gain for each inducing point.

The Gaussian likelihood treats the rehearsal targets at different inducing points as conditionally independent given their function values. Its expected log-likelihood therefore depends on $\mathbf{S}$ only through the marginal variances S_{mm}. Weighting Eq. S12 by the attention gains gives $\sum_m a(\ell_m^z) S_{mm} = \operatorname{tr}(\mathbf{AS})$ and

$$-\frac{1}{M}\sum_{m=1}^{M} a(\ell_m^z)\,\mathbb{E}_q\big[\log p(t_m \mid f(\mathbf{z}_m))\big] = \frac{1}{2\sigma^2 M}(\mathbf{m}-\mathbf{t})^\top \mathbf{A}(\mathbf{m}-\mathbf{t}) + \frac{1}{2\sigma^2 M}\operatorname{tr}(\mathbf{AS}) + \text{const.} \tag{S19}$$

The KL term of Eq. 5 is β times Eq. S13,

$$\beta\,\mathrm{KL}\big(q(\mathbf{u})\,\|\,p(\mathbf{u})\big) = \frac{\beta}{2}(\mathbf{m}-\mu_0\mathbf{1})^\top \mathbf{K}_{ZZ}^{-1}(\mathbf{m}-\mu_0\mathbf{1}) + \frac{\beta}{2}\operatorname{tr}(\mathbf{K}_{ZZ}^{-1}\mathbf{S}) - \frac{\beta}{2}\log|\mathbf{S}| + \text{const.} \tag{S20}$$

With the rehearsal target $\mathbf{t}$ held fixed, the sum of (S19) and (S20) separates into terms involving the mean $\mathbf{m}$ and terms involving the covariance $\mathbf{S}$. We can therefore differentiate the objective with respect to each quantity separately. The mean-dependent terms are

$$\mathcal{L}_{\text{maint}}(\mathbf{m}) = \frac{1}{2\sigma^2 M}(\mathbf{m}-\mathbf{t})^\top \mathbf{A}(\mathbf{m}-\mathbf{t}) + \frac{\beta}{2}(\mathbf{m}-\mu_0\mathbf{1})^\top \mathbf{K}_{ZZ}^{-1}(\mathbf{m}-\mu_0\mathbf{1}) + \text{const.} \tag{S21}$$

Differentiating the mean-dependent objective with respect to $\mathbf{m}$ gives

$$\nabla_{\mathbf{m}}\mathcal{L}_{\text{maint}} = \frac{1}{\sigma^2 M}\mathbf{A}(\mathbf{m}-\mathbf{t}) + \beta\,\mathbf{K}_{ZZ}^{-1}(\mathbf{m}-\mu_0\mathbf{1}) = \mathbf{W}\mathbf{m} - \mathbf{b}, \tag{S22}$$

Updating the mean in the direction of the negative gradient gives the recurrent equation:

$$\tau\,\dot{\mathbf{m}} = -\mathbf{W}\mathbf{m} + \mathbf{b}, \qquad \mathbf{W} = \frac{1}{\sigma^2 M}\mathbf{A} + \beta\,\mathbf{K}_{ZZ}^{-1}, \qquad \mathbf{b} = \frac{1}{\sigma^2 M}\mathbf{A}\,\mathbf{t} + \beta\,\mathbf{K}_{ZZ}^{-1}\mu_0\mathbf{1}. \tag{S23}$$

Here $\mathbf{A} = \operatorname{diag} a(\ell_m^z)$ is the diagonal gain matrix from Eq. 7; $\mathbf{W}$ determines recurrent interactions, and $\mathbf{b}$ is the drive from rehearsal targets and the prior mean. The numerical implementation uses Adam in whitened coordinates and refreshes $\mathbf{t} \leftarrow \mathbb{E}_q[f(\mathbf{Z})]$ after each step (S3.3). The derivative (S22) treats $\mathbf{t}$ as constant.

Supplementary Fig. S3 shows how the prior-precision term of $\mathbf{W}$ depends on the kernel. All panels use the 10×10 inducing grid of Fig. 5b and show $-(\mathbf{K}_{ZZ} + \varepsilon\mathbf{I})^{-1}$ from one unit, with the self-coupling removed and weights scaled by their largest absolute entry. For a product kernel $k(\ell, \ell')\,k(c, c')$ on the grid, $\mathbf{K}_{ZZ}$ is a Kronecker product, so its inverse couples diagonal neighbors through the product of two one-dimensional nearest-neighbor terms. The periodic kernel with $\lambda = 10$ is nearly singular on this grid: the smallest eigenvalue of its one-dimensional 10×10 kernel matrix is 5×10^{-8}. Modes with eigenvalues below ε are therefore set by the jitter, and the weight pattern changes with ε.

The covariance-dependent terms do not depend on the rehearsal target $\mathbf{t}$. Adding the variance term of (S19) to the terms in (S20) that depend on $\mathbf{S}$ gives

$$\begin{aligned}
\mathcal{L}_{\text{maint}}(\mathbf{S}) &= \frac{1}{2\sigma^2 M}\operatorname{tr}(\mathbf{A}\mathbf{S}) + \frac{\beta}{2}\operatorname{tr}(\mathbf{K}_{ZZ}^{-1}\mathbf{S}) - \frac{\beta}{2}\log|\mathbf{S}| \\
&= \frac{1}{2}\operatorname{tr}\left(\left(\frac{1}{\sigma^2 M}\mathbf{A} + \beta\,\mathbf{K}_{ZZ}^{-1}\right)\mathbf{S}\right) - \frac{\beta}{2}\log|\mathbf{S}| \\
&= \frac{1}{2}\operatorname{tr}(\mathbf{W}\mathbf{S}) - \frac{\beta}{2}\log|\mathbf{S}|,
\end{aligned} \tag{S24}$$

with the same $\mathbf{W}$ as in (S23). Encoding and maintenance have related quadratic forms, but their likelihood projections, normalization factors and KL coefficients differ. For symmetric $\mathbf{W}$ and $\mathbf{S}$, the derivative is

$$\frac{\partial \mathcal{L}_{\text{maint}}}{\partial \mathbf{S}} = \frac{1}{2}\mathbf{W} - \frac{\beta}{2}\mathbf{S}^{-1}. \tag{S25}$$

Setting the covariance gradient to zero yields the stationary covariance $\mathbf{S}^* = \beta\,\mathbf{W}^{-1}$. Gradient descent on (S24) follows this derivative. The common factor of one half in the gradient can be absorbed into the time constant τ. The resulting continuous-time gradient update is

$$\tau\,\dot{\mathbf{S}} = -(\mathbf{W} - \beta\,\mathbf{S}^{-1}), \qquad \mathbf{S}^* = \beta\,\mathbf{W}^{-1}. \tag{S26}$$

Increasing the diagonal entries of $\mathbf{A}$ near the cue increases the corresponding diagonal entries of $\mathbf{W}$ and reduces the corresponding stationary marginal variances in $\mathbf{S}^*$ when the other quantities are fixed. The simulated finite delay need not reach this stationary covariance. To ensure a valid covariance during numerical optimization, we use the factorization $\mathbf{S} = \operatorname{tril}(\mathbf{C})\operatorname{tril}(\mathbf{C})^\top$, where tril retains the lower-triangular part of its matrix argument. Multiplying the factor by its transpose

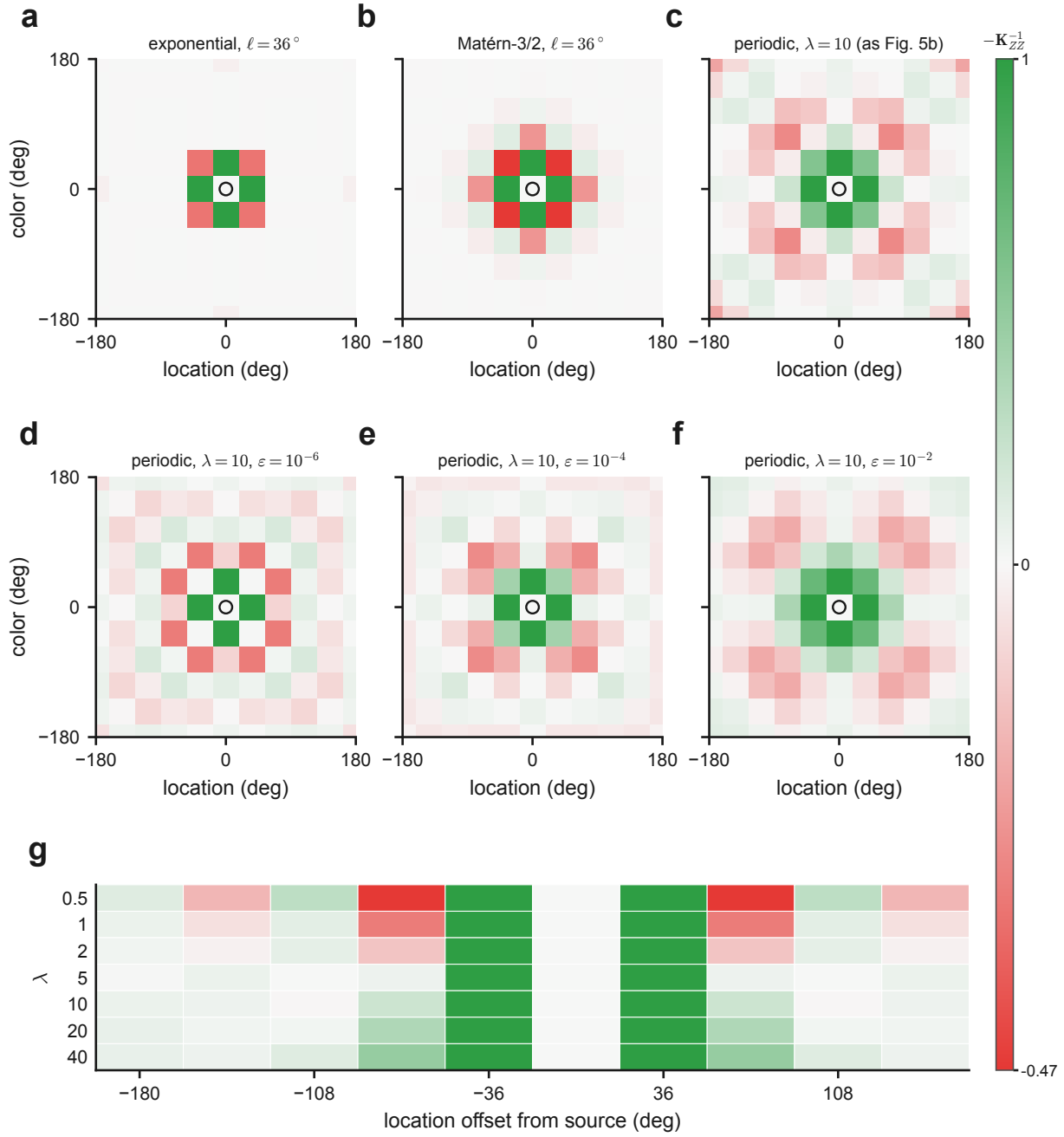


Figure S3: Kernel smoothness and conditioning shape the derived connectivity. (a–c) Outgoing weights $-(\mathbf{K}_{ZZ} + \varepsilon\mathbf{I})^{-1}$ from the unit at the origin (open circle; self-coupling removed) on the 10×10 inducing grid with $\varepsilon = 10^{-3}$, for an exponential kernel (a) and a Matérn-3/2 kernel (b), both with lengthscale 36° (one grid step), and for the periodic kernel with $\lambda = 10$ used in the paper (c; as in Fig. 5b). Each is a product kernel over location and color. Weights are scaled by their largest absolute entry; green and red indicate excitation and inhibition. **(d–f)** The periodic kernel with $\lambda = 10$ at $\varepsilon = 10^{-6}$ (d), 10^{-4} (e) and 10^{-2} (f). **(g)** Weights along the location axis at the source color for the periodic kernel with λ from 0.5 to 40 ($\varepsilon = 10^{-3}$).

ensures positive semidefiniteness. The implementation applies this factorization to the whitened covariance $\mathbf{S}_v$ and updates its factor with Adam (S3.3). Gradient flow in $\mathbf{S}$ and Adam updates in its Cholesky factor optimize related objectives but do not have identical trajectories.

S3.3 Numerical coordinates and circuit comparison

The derivation above uses the original inducing values, but the numerical implementation uses a transformed parameterization. GPyTorch computes variational updates in whitened coordinates: a linear transformation centers the inducing variables and gives them identity prior covariance. If $\mathbf{L}_K\mathbf{L}_K^\top = \mathbf{K}_{ZZ}$, the unwhitened moments are $\mathbf{m} = \mu_0\mathbf{1} + \mathbf{L}_K\mathbf{m}_v$ and $\mathbf{S} = \mathbf{L}_K\mathbf{S}_v\mathbf{L}_K^\top$, where $\mathbf{m}_v$ and $\mathbf{S}_v$ are the whitened moments. When constructing the quadratic operator, the numerical implementation uses the same predictive evaluation and jitter as GPyTorch's variational strategy. Jitter is a small diagonal covariance term used for numerical stability. The analytic equations above express the model in unwhitened coordinates, which expose the prior-dependent lateral interactions.

For the recurrent-circuit maintenance simulation, the operator and rehearsal-dependent drive are constructed from derivatives of the fixed-geometry GP objective with respect to the whitened variational mean. Mean updates and updates of the whitened Cholesky covariance then use Adam with the encoding-end first moments, second moments and step counts. The rehearsal target is refreshed after each update. At cue onset, the attention gains change the diagonal likelihood contribution to the recurrent operator. This procedure compares the circuit and MemGP under the same numerical update convention; Adam is not invariant to changing between whitened and unwhitened coordinates. In this comparison, MemGP additionally permits the inducing locations, prior mean and kernel parameters to change during maintenance.

The equilibrium results in S3.1–S3.2 assume that the prior, inducing locations and rehearsal targets remain fixed, in which case the objective has a unique minimum with respect to the mean. In the maintenance simulation, the rehearsal target is updated after each step, so the input drive changes across steps and these equilibrium results apply only approximately to a simulated delay. In the ideal self-consistent limit $\mathbf{t} = \mathbf{m}$, the rehearsal contribution to the instantaneous mean gradient vanishes, leaving the prior-dependent term; attention still enters the covariance equation. Attention can therefore affect the covariance even when its direct contribution to the instantaneous mean gradient vanishes. The mean and covariance need not show the same response to a cue.

S3.4 Retrieval

The maintained mean and covariance $(\mathbf{m}, \mathbf{S})$ define a posterior over memory strengths. The circuit uses the same retrieval rule as the behavioral simulations: draw strengths jointly across colors and occupied locations, subtract the global sampled minimum to obtain nonnegative responses, and apply spatial divisive normalization as specified in S1.5. The report is $\hat{c} = \arg\max_c r(c)$. Sampling and normalization are specified readout operations; they are not derived as additional recurrent dynamics from the variational objective.