bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A dynamic neural resource model bridges sensory and working memory

Ivan Tomić[1,2,*] & Paul M. Bays[1]

[1]University of Cambridge, Department of Psychology, Cambridge, UK
[2]University of Zagreb, Department of Psychology, Zagreb, CRO
[*]corresponding author: ivn.tomic@gmail.com

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Probing memory of a complex visual image within a few hundred milliseconds after its disappearance reveals significantly greater fidelity of recall than if the probe is delayed by as little as a second. Classically interpreted, the former taps into a detailed but rapidly decaying visual sensory or "iconic" memory (IM), while the latter relies on capacity-limited but comparatively stable visual working memory (VWM). While iconic decay and VWM capacity have been extensively studied independently, currently no single framework quantitatively accounts for the dynamics of memory fidelity over these timescales. Here we extend a stationary neural population model of VWM with a temporal dimension, incorporating rapid sensory-driven accumulation of activity encoding each visual feature in memory, and a slower accumulation of internal error that causes memorized features to randomly drift over time. Instead of facilitating read-out from an independent sensory store, an early cue benefits recall by lifting the effective limit on VWM signal strength imposed when multiple items compete for representation, allowing memory for the cued item to be supplemented with information from the decaying sensory trace. Empirical measurements of human recall dynamics validate these predictions while excluding alternative model architectures.

Keywords: short-term memory, population coding, temporal dynamics, delay, encoding, decoding

Significance

The need to make sense of and interact with the world often requires us to keep information from our senses in mind for short periods of time. This ability is constrained by how quickly the brain can incorporate new sensory information into short-term memory, the limited capacity of that memory and the rate at which memories deteriorate. Here we propose a new mechanistic account, based on principles of neural coding, that unifies processes of encoding, sensory and working memory in a comprehensive framework that captures temporal dynamics in the fidelity of human short-term recall. A key conclusion is that sensory information cannot contribute directly to a cognitive judgment, but must first be integrated into resource-limited working memory.

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Keeping relevant information in an easily accessible state is vital for adaptive behavior in dynamic environments. In the primate visual system, this requirement is met by visual working memory (VWM), the capacity to actively maintain visual information from milliseconds to seconds after a stimulus disappears from view [1–4]. While the contents of VWM are frequently updated to reflect changes in the environment and in behavioral priorities, the visual processing hierarchy itself introduces additional layers of dynamism [5, 6]. The fidelity of representations therefore evolves from the moment VWM starts accumulating evidence [7, 8] throughout the maintenance period until the information is used for action [9–11].

Nonetheless, within most theoretical frameworks, VWM is treated as a stationary process whereby representations are measured and modeled as fixed states of the system. One such model of working memory is based on principles of neural population coding [12, 13]. In the Neural Resource model, visual information is encoded in the activity of a population of noisy feature-selective neurons [14, 15]. The spiking activity of the neural population is constrained by normalization [16], such that the total activity is fixed but flexibly distributed between memoranda, implementing a form of limited memory resource. At retrieval, encoded stimulus values are reconstructed from the noisy spiking activity. This model has provided a quantitative account of patterns of recall error across a range of tasks and stimulus dimensions [17–20]. However, despite its grounding in principles of neural coding, the basic architecture of the model lacks a temporal dimension to describe the dynamics of memory representations during encoding and maintenance.

Research on prolonged memory maintenance has demonstrated that the precision of stored representations gradually deteriorates over time (e.g., 21, 22). Computational models attempting to account for these dynamics have often relied on principles of diffusion within an attractor network. In such a network, information is maintained in a sustained pattern of activity, which can be visualized as a "bump" of activity centered on the stored value. Over time, the bump diffuses along the feature dimension due to random fluctuations in neural activity, leading to stochastic changes in the encoded feature value and a gradual loss of information [23, 24]. Critically, the neural code diffuses without decay in signal strength. A growing body of empirical support, both at the behavioral [9] and neural level [25, 26], identifies diffusion as a key mechanism of memory deterioration.

In contrast to such gradual deterioration over longer retention intervals, studies that probed memory within a few hundred milliseconds of stimulus offset revealed a precipitous decrease in memory fidelity immediately after a stimulus disappears [27–30]. This early superior recall was attributed to a high-capacity but short-lived form of storage termed iconic memory (IM) [31]. The behavioral advantage of early cues has been ascribed to reading out information directly from IM and circumventing capacity limitations imposed by VWM, however, this idea has not been formally modelled or tested. At the neural level, IM is thought to be supported by a brief period of decaying neural activity in early visual areas following the response elicited by the visible stimulus [32–34]. In contrast to later memory dynamics arising due to noise accumulation, early changes in memory fidelity were supported by modulation of the neural signal strength. However, little is known about the read-out of this sensory memory buffer.

Finally, memory fidelity changes during encoding while the evidence is extracted from the visible stimulus. Previous studies revealed that longer stimulus exposures have a favorable effect on the subsequent recall, but that this effect is modulated by the number of simultaneously encoded objects [35–37], providing evidence for a processing or encoding limitation of VWM. As stimulus presentation duration increases, more information may be extracted from the sensory signal into VWM, increasing the fidelity of the representation. Critically, with prolonged exposure, VWM fidelity approaches a stable level that depends on the number of encoded items, suggesting that a ceiling is imposed on evidence accumulation by a shared limit on VWM resources. However, a computational framework describing information accumulation from sensory areas into VWM is lacking, and the observed encoding limit may reflect dynamics in sensory areas registering visible objects as well as VWM accumulating this sensory evidence.

Here, we investigated the temporal dynamics in the fidelity of VWM from information encoding

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

until its recall. To map human recall fidelity to the time domain, we conducted psychophysical experiments in which we probed memory representations at different time points relative to stimulus onset and offset while simultaneously manipulating set size. To isolate memory dynamics due to changes in the representational signal, we advanced an analogue reproduction task with a novel response method specifically adapted to minimize the time cost of motor (i.e., response) processes and capture the momentary state of memory representations. This allowed us to precisely measure the time course of fidelity dynamics during representation formation (i.e., encoding) and retention (i.e., maintenance). A major conclusion is that the enhanced precision seen at very brief retention intervals depends on integration of information from the sensory store into VWM following the cue, with direct read-out from IM unable to account for the empirical patterns of results.

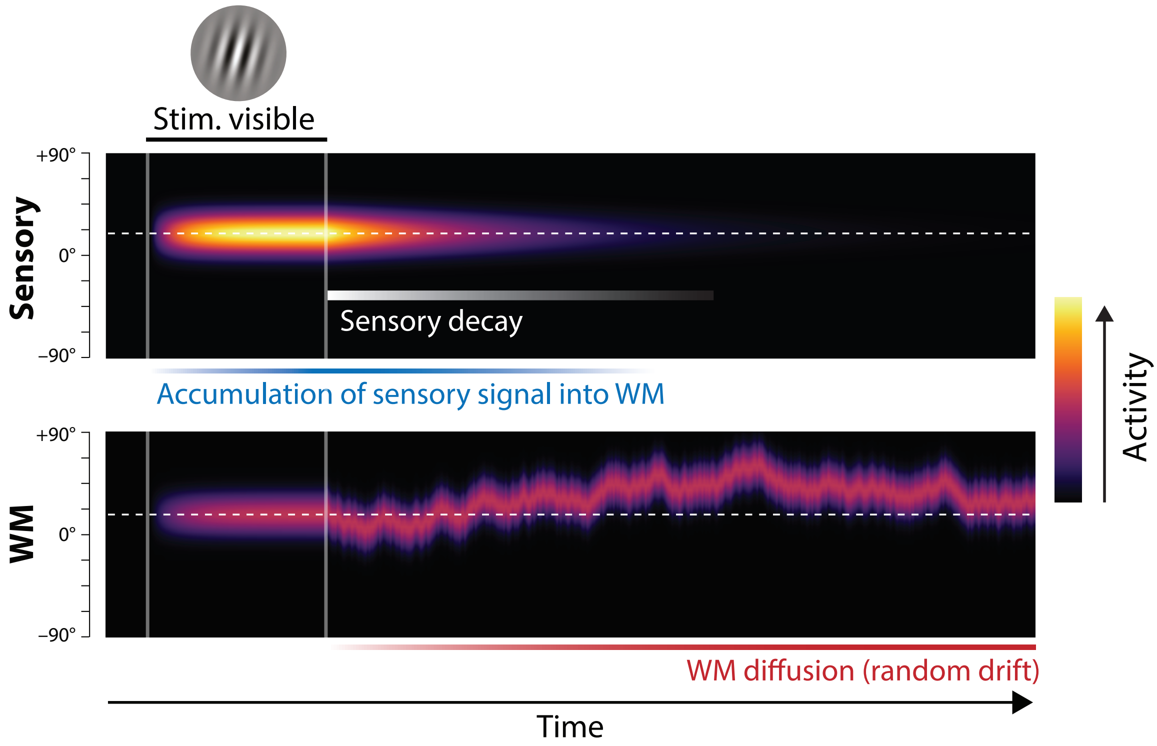

Figure 1: Proposed neural population dynamics for encoding a single orientation into VWM and maintaining it over a delay. Top: Stimulus onset is followed by a ramping increase in activity (indicated by color) of sensory neurons whose tuning (indicated on y axis) matches the stimulus orientation. Following stimulus offset, this sensory signal rapidly decays. The sensory signal, including its decaying post-stimulus component, provides input into VWM. Bottom: At stimulus onset, the VWM population begins to accumulate activity from the sensory population. This accumulation saturates at a maximum amplitude determined by global normalization. As the sensory activity decays, the activity in the VWM population is maintained at a constant amplitude, but accumulation of random errors causes the activity bump to diffuse along the feature dimension (y axis) over time, changing the orientation represented by the population. At recall, when the VWM population activity is decoded, accuracy of the recall estimate depends on both the orientation represented (centre of the activity bump) and the fidelity with which it can be retrieved (determined by activity amplitude).

To explain the neural computations underlying the observed time courses, we devised a comprehensive neural model of memory dynamics whose core architecture is rooted in the Neural Resource model of VWM [12, 13]. The Dynamic Neural Resource (DyNR) model assumes that changes in memory fidelity reflect temporal dynamics in the sensory population registering the stimuli and from signal and noise accumulation processes of resource-limited VWM (Fig. 1). In particular, the model prescribes how time-dependent gain control mechanisms in sensory areas produce a smooth neural response following abrupt changes in stimulus presence. As this sensory signal provides feed-forward input to VWM, the dynamics in VWM activity in the temporal vicinity of stimulus presentation (i.e., onset and offset) strongly reflect not only limits in VWM, but also the dynamics of the sensory

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

signal. Finally, once accumulated into VWM, the neural signal is subject to perturbations due to noise accumulation, resulting in degradation of internal representations with time. The DyNR model accurately reproduced the detailed empirical patterns of human recall errors in the psychophysical experiments. Based on these results, we argue that changes in memory fidelity on short time scales reflect dynamics in the gain or signal strength in neural populations representing the stimulus, while changes on longer time scales are dominated by corruption of the representation by accumulated noise.

Dynamic Neural Resource (DyNR) Model

The Dynamic Neural Resource model generalizes an established neural population account of VWM, originally proposed by Bays [12] and inspired by similar models of attention and perceptual decision-making [38, 39]. In the original model, memorization and recall of visual stimuli is achieved by encoding and decoding of spiking activity in idealized feature-tuned neurons. The limited capacity of VWM to hold multiple object features simultaneously is reproduced by a global divisive normalization that constrains total spiking activity, implementing a continuous memory resource [16, 12]. The DyNR model (illustrated in Fig. 1) extends this stationary encoding-decoding model with a temporal dimension. First, to capture encoding dynamics, stimulus information enters the VWM population (Fig. 1, bottom) indirectly, by accumulation of neural signal from a separate sensory population (top), which receives the visual input. The signal strength in the VWM population at any point in time jointly depends on the history of the signal in the sensory population and the number of features competing for representation in VWM. Once the sensory signal is gone, the VWM signal is maintained at its maximum attained amplitude, but the stimulus value encoded by the signal gradually diffuses due to accumulation of random noise. Recall error depends on both the stimulus value represented at the time of retrieval (*what* is encoded) and the signal amplitude at that time, read out in the form of spikes (*how precisely* it can be decoded).

Dynamics of sensory signal strength

To model the temporal dynamics of human memory fidelity, we begin by defining computations of the sensory system registering the incoming signal. A particularly important computation is temporal filtering – a property of neurons to respond more sensitively to specific temporal patterns in stimuli. To model the signal represented in the cortical sensory level, we assume that the sensory response to a stimulus presentation of fixed duration (described as a step function in visual input amplitude, Fig. 2A & B, left) is controlled by a monophasic temporal filter having a low-pass frequency response [40]. This choice is a natural one since it is consistent with electrophysiological studies demonstrating that a large range of temporal frequencies registered by the retina and LGN [41, 42] is attenuated at higher frequencies before the signal enters the primary visual cortex [43]. Passing the stimulus through such a temporal filter attenuates the neural response to fast transients in the signal, and thereby produces a smooth rise and decay of neural activity in response to a uniform input signal (Fig. 2C). In particular, we assume that the activity of the sensory population after stimuli onset and offset changes exponentially towards the maximum sensory activity and baseline activity, respectively.

The choice of the filter's temporal response characteristics (i.e., its time constant) fully defines dynamics in the sensory population activity and controls the signal projected towards higher areas. The available physiological evidence suggests the temporal properties of the rising and decaying neural response are not symmetric [44, 45]. In particular, the neural response typically reaches the maximum activity after the onset faster than it reaches the baseline activity after the offset. Consistent with this, we allowed the sensory signal to decay at a different rate than the rising rate. The temporal dynamics in sensory population firing activity in response to a fixed input signal of duration t_{offset} is then given by:

$$\dot{\gamma}_{\mathrm{s}}(t) = \begin{cases} (\check{\gamma}_{\mathrm{s}} - \gamma_{\mathrm{s}}(t))/\tau_{\mathrm{rise}} & \text{for} \quad t \leq t_{\mathrm{offset}} \\ -\gamma_{\mathrm{s}}(t)/\tau_{\mathrm{decay}} & \text{for} \quad t > t_{\mathrm{offset}} \end{cases} \tag{1}$$

where $\check{\gamma}_{\mathrm{s}}$ is the maximum sensory signal, τ_{rise} and τ_{decay} are rising and decaying time constants of the temporal filter, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The temporal properties of the sensory response have been shown to depend on the physical characteristics of stimuli, such as contrast and location [44, 46]. Similarly, previous work has demonstrated that the decaying component of the sensory response is strongly influenced by the engagement of the sensory population after stimuli offset (e.g., 32). In particular, a new input signal, e.g. a backward noise mask, curtails ongoing activity related to the previous stimulus, resulting in a faster decay of activity compared to the unmasked post-stimulus period [47]. Consistent with this, here we assume that the backward mask operates by interrupting ongoing sensory processing of stimuli, limiting the access to the sensory signal (cf. integration mask) [48].

Dynamics of VWM signal strength

The information registered by the sensory system is subsequently accumulated into a VWM population capable of maintaining activity in the absence of further input (e.g. by self-excitation, see 49, 50, 24; although only the resulting dynamics are modelled here). The total activity of the VWM neural population is normalized, implementing a limited resource shared out between memory items [12, 13]. Consequently, if the stimuli are presented for long enough, the evidence accumulated from the sensory signal into VWM will saturate at a level that reflects the total number of stimuli represented (Fig. 2D). The dynamics in VWM population activity are given by:

$$\dot{\gamma}_{\mathrm{wm}}(t) = \gamma_{\mathrm{s}}(t)(\check{\gamma}_{\mathrm{wm}}/M(t) - \gamma_{\mathrm{wm}}(t))/\tau_{\mathrm{wm}} \tag{2}$$

where $\check{\gamma}_{\mathrm{wm}}$ is the maximum VWM signal amplitude, $M(\mathrm{t})$ is the number of items represented in VWM at time t, τ_{wm} is the time constant of accumulation into VWM.

A common assumption of VWM models is that the strength of the representational signal remains stable after encoding from a visible stimulus. This stationary view has been reinforced by typically measuring VWM sufficiently long after the stimulus disappears ($\sim$1 second) and at a single time-point. In contrast, work on IM demonstrated that recall fidelity in a brief period after stimulus offset typically surpasses and then precipitously decays towards VWM fidelity level [51]. Consistent with that, we consider how the normalized representational signal in VWM formed during encoding can be boosted in the absence of the physical stimulus. In particular, we assume a representation stored in VWM can be strengthened as long as the sensory population provides feed-forward input and VWM activity is not saturated at the normalized level. Such a scenario can be achieved by cueing an item for recall in the temporal vicinity of stimulus offset, i.e. before sensory activity decays to zero. By cueing an item for recall, the remaining contents of VWM becomes obsolete and can be removed from memory [52]. In the model,

$$M(t) = \begin{cases} N & \text{for} \quad t \leq t_{\mathrm{cue}^*} \\ 1 & \text{for} \quad t > t_{\mathrm{cue}^*} \end{cases} \tag{3}$$

where t_{cue^*} is the time when the item is identified for a recall and the readout of stimulus value begins. This "demounting" of resource from uncued items makes it available for storing additional information about the cued item, which is extracted from the residual sensory representation, increasing the representation fidelity beyond that granted by equal distribution of neural signal between items. Critically, as sensory information quickly decays, there will be less signal remaining to supplement the VWM representation of a cued item if the cue is delivered later, and at the longest cue intervals the cue will confer no advantage over the fidelity attained when all items compete equally for VWM representation (Fig. 2D).

We note that removal of uncued items cannot occur until the cue has been processed to the point of identifying one of the N items in the memory array. We follow Hick [53] in modelling this cue processing time as logarithmic in the number of alternatives:

$$t_{\mathrm{cue}^*} = t_{\mathrm{cue}} + b \log_2(N) \tag{4}$$

where b is a scaling parameter. Previous work demonstrated that estimation of temporal dynamics in attention and memory could be confounded with the time needed to interpret the cue and start acting on it [54]. This is especially significant when trying to accurately capture quickly changing processes, such as decay of the sensory residual. Although the cue processing time likely fluctuates on a trial-by-trial basis due to changes in, e.g. attention, arousal, or motivation, here we focus on the influence of set size arising from a limited information processing capacity.

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Diffusion of VWM encoded values

So far we have described only changes in the strength of the neural signal encoding features in memory. However, feature representations maintained over time in neural activity will accumulate noise in the absence of external input. We model this process of noise-driven diffusion as Brownian motion in feature space throughout the retention interval (Fig. 1), contributing to variability in the decoded feature value [23, 9]. The resulting variability is described by a wrapped normal distribution with variance σ^2 that increases linearly with time from stimulus offset, so that at time t the encoded feature corresponding to a true stimulus feature θ is

$$\theta(t) \sim \mathcal{WN}(\theta, \sigma^2(t)) \tag{5}$$

$$\sigma^2(t) = (\mathrm{t} - \mathrm{t}_{\mathrm{offset}})\dot{\sigma}^2_{\mathrm{diff}} \tag{6}$$

where $\dot{\sigma}^2_{\mathrm{diff}}$ specifies the base diffusion rate. While the fast decay of sensory activity after stimuli offset accounts for early dynamics in VWM fidelity, diffusion becomes prominent over longer delays, accounting for more gradual deterioration of precision with time.

Such a diffusion account has support in the available neural evidence as well as in theoretical work. At the neural level, an electrophysiological study in monkeys performing a spatial working memory task demonstrated that shifts of neural tuning curves during a memory delay predicted behavioral response errors [24]. A similar finding was observed in humans where drift in the fMRI activity patterns relative to the target predicted errors in an orientation discrimination task [25]. At a theoretical level, continuous attractor models explain diffusion as a consequence of neural variability in networks where excitatory and inhibitory connections constrain population activity to a sub-space or manifold corresponding to the encoded feature space [23, 55, 50].

Retrieval

To model the process that leads to a response we first consider that in some trials observers may erroneously identify a non-target item as being cued. Previous work indicates these "swap" errors occur due to uncertainty in memory for the cue features of the stimuli, in this case their locations [19, 56]. We assume that changes in variability in the cue features mirror those of the memory features, leading swap frequency to decrease exponentially as a function of presentation duration and increase linearly with retention interval (Fig. S3):

$$p_{\mathrm{swap}} = (N-1)\left[\left(\frac{1}{N} - r_{\mathrm{spatial}}t_{\mathrm{cue}^*}\right)e^{\frac{-t_{\mathrm{offset}}}{\tau_{\mathrm{spatial}}}} + r_{\mathrm{spatial}}t_{\mathrm{cue}^*}\right] \tag{7}$$

where τ_{spatial} is the time constant related to presentation duration, and r_{spatial} is the rate constant related to the retention interval.

If θ is the true feature value of the item identified as the target (i.e. the cued item with probability $1 - p_{\mathrm{swap}}$, a randomly selected non-cued item with probability p_{swap}), then due to diffusion (Eq. 5) the value encoded in the VWM population at the time of retrieval is given by

$$\theta^* \sim \mathcal{WN}(\theta, \sigma^2(t_{cue^*})) \tag{8}$$

We model retrieval as estimation of θ^* based on spiking activity in the VWM population that encodes the selected item. For this purpose we assume an idealized set of tuning functions, where the mean response of neuron i encoding orientation θ with population gain γ is described by

$$f_i(\theta, \gamma) = \frac{\gamma}{n}\exp(\kappa(\cos(\theta - \varphi_i) - 1) \tag{9}$$

where n is the number of neurons, and κ determines the tuning width. The preferred orientations of the neurons, φ_i, are evenly distributed throughout the circular space to provide uniform coverage. The spike count produced by each neuron is drawn from a Poisson distribution,

$$r_i \sim \mathrm{Poisson}(f_i(\theta^*, \gamma_{wm^*})) \tag{10}$$

and the decoded orientation estimate is obtained by maximum likelihood estimation based on the spike counts:

$$\hat{\theta} = \arg\max_{\theta} p(\mathbf{r}|\theta). \tag{11}$$

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Additional assumptions

To fit the model to behavioral data, we make several further simplifying assumptions. We assume that the exponential decay of the sensory signal is rapid enough that there is effectively no information remaining by the time the VWM population is decoded to generate a response. This allows us to approximate the VWM activity at the time of decoding by the asymptotic VWM activity were the sensory decay to continue indefinitely:

$$\gamma_{\mathrm{wm}^*} \approx \gamma_{\mathrm{wm}}(\infty) \tag{12}$$

Next, we identify diffusion in the encoded value at the time of retrieval with diffusion at the time of target item identification (justifying the use of t_{cue^*} in Eq. 8. We reason that the rate of diffusion is slow enough relative to the rate of sensory decay, that any additional diffusion in the brief period of post-cue sensory accumulation is negligible.

In Experiment 1 (see below), a task with a fixed 200 ms exposure period, we assume that the initial encoding of all items into VWM is complete by the time of stimulus offset, i.e. that VWM activity at this time can be approximated by its asymptotic level reflecting normalization:

$$\gamma_{\mathrm{wm}}(t_{\mathrm{offset}}) \approx \check{\gamma}_{\mathrm{wm}}/N \tag{13}$$

Finally, in the condition of Experiment 1 where memory array and cue are presented simultaneously, we assume that only the cued feature is encoded in VWM, reaching the maximum amplitude, $\check{\gamma}_{\mathrm{wm}}$, irrespective of set size. Maximum likelihood fits were obtained via the Nelder-Mead simplex method (function *fminsearch* in Matlab). All parameters and variables used to describe the DyNR model are listed in Table 1.

Table 1: DyNR model parameters (1–9) and other variables (10–24) used in model description.

No.	Parameter/variable	Description
1	$\check{\gamma}_{\mathrm{wm}}$	Maximum VWM signal amplitude
2	κ	Tuning curve width
3	τ_{rise}	Rise constant of the sensory temporal filter
4	τ_{decay}	Decay constant of the sensory temporal filter
5	τ_{wm}	Time constant of accumulation into VWM
6	$\dot{\sigma}^2_{\mathrm{diff}}$	Base diffusion rate
7	τ_{spatial}	Time constant for spatial encoding
8	r_{spatial}	Rate constant for spatial diffusion
9	b	Scaling parameter for Hick's law
10	t	Time, relative to stimulus onset ($t = 0$)
11	t_{offset}	Time of stimulus offset
12	t_{cue}	Time of cue onset
13	t_{cue^*}	Time an item is identified for report
14	N	Number of items in stimulus array
15	$M(t)$	Number of items in memory at time t
16	$\check{\gamma}_{\mathrm{s}}$	Maximum sensory signal amplitude
17	$\gamma_{\mathrm{s}}(t)$	Sensory signal amplitude at time t
18	$\gamma_{\mathrm{wm}}(t)$	VWM signal amplitude at time t
19	γ_{wm^*}	VWM signal amplitude at the time of decoding
20	$\sigma^2(t)$	Accumulated diffusion at time t
21	n	Number of neurons
22	θ	True stimulus feature value
23	θ^*	Encoded stimulus feature value at the time of decoding
24	$\hat{\theta}$	Decoded stimulus feature value

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

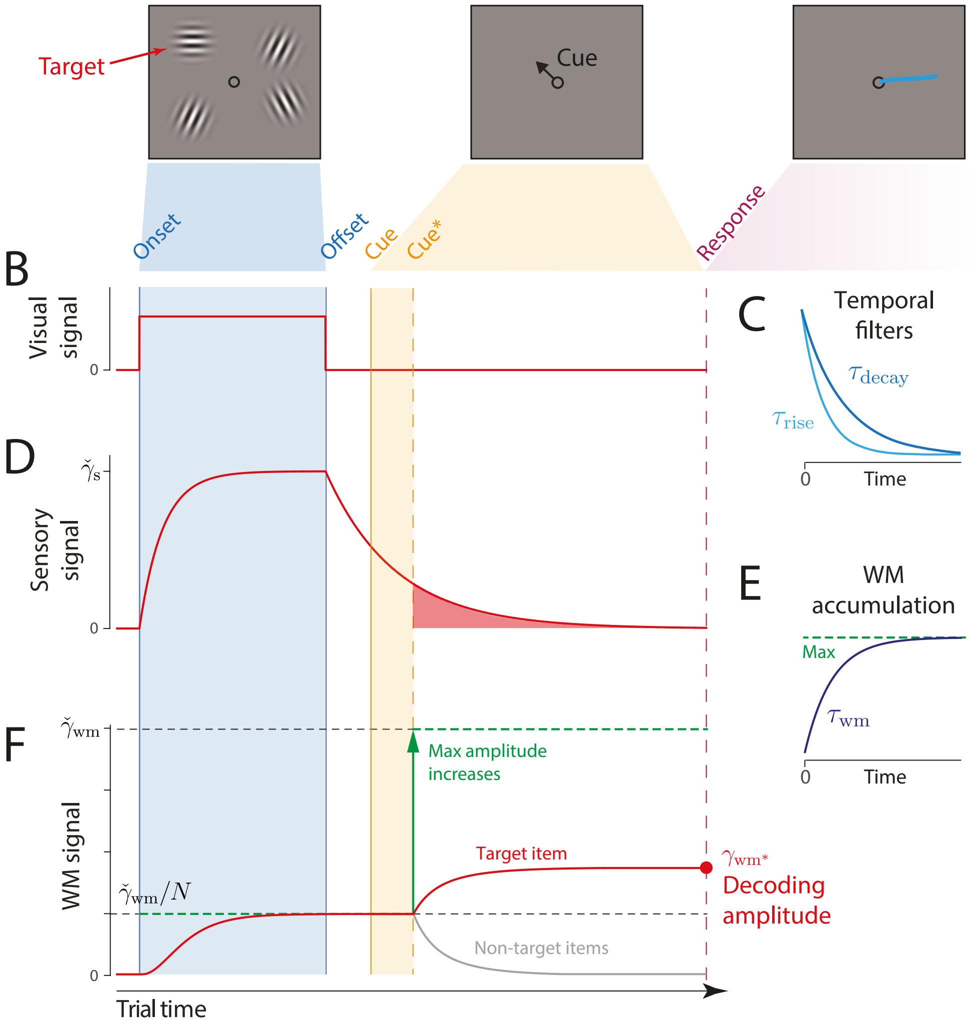

Figure 2: Schematic of signal amplitudes in the DyNR model during a cued recall trial. (**A**) Observers are presented with a memory array (left), followed after a blank delay (not shown) by an arrow cue (center) indicating the location of one item (the target) whose remembered orientation should immediately be reported (right). (**B**) The amplitude of the visual input associated with each item is modelled as a step function (left). The sensory response (**D**) is modelled as a low-pass filtering of the stimulus input, with different time constants for rise and decay (**C**). (**F**) Amplitude of the working memory signal reflects a saturating accumulation of activity from the sensory population (illustrated in **E**). Beginning with stimulus onset, activity associated with each item is accumulated from the sensory population into the VWM population, approaching an upper bound (green dashed line) that reflects a total activity limit shared between the N items in memory. Once the cue has been presented (solid orange line) and processed (dashed orange line), uncued items can be dropped from VWM, raising the ceiling on activity available to represent the cued item (green arrow). This allows more information about the cued item to be accumulated from the decaying sensory trace (equivalent to the red shaded area in D). Response variability depends on the asymptotic VWM signal amplitude available for decoding (red circle) combined with the accumulated effects of diffusion (see text).

Overview of Experiments

We tested predictions of the Dynamic Neural Resource model against empirical data collected in continuous report tasks. In Experiment 1 (Fig. S1A & B), observers were presented with an array of oriented stimuli for a fixed duration followed after a variable delay by a visual cue identifying one of the preceding stimuli whose orientation should be reported. This experiment was designed to investigate the contribution of decaying sensory representations following stimulus offset to the

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

dynamics of recall fidelity. Experiment 2 (Fig. S1C) was aimed at expanding the results of the first experiment to now also assess the accumulation of information during the time the stimuli were visible. In this case, the exposure duration was varied while the delay before the visual cue was held constant. In both experiments we varied the number of stimuli in the array (set size) to assess capacity limitations affecting encoding and maintenance.

To provide additional validation of the DyNR model, we also tested its predictions against data from a previously published continuous report experiment (Experiment 1 in 12) and one additional dataset collected as part of a separate study [57]. A detailed description of all experiments is provided in Supplementary Information.

Results

Experiment 1: Delay duration

In Experiment 1, we evaluated the time course of VWM fidelity over brief memory intervals. Previous work has demonstrated that immediately after a stimulus physically disappears, its representation briefly persists in the sensory system in the form of residual neural activity [33]. Accumulation of this lingering sensory activity into VWM could enable superior recall of information [51] within the constraints of a finite VWM resource that strongly limits representational fidelity [3]. To describe these dynamics, we examined human recall of orientation stimuli presented in arrays of varying sizes and probed after a variable delay ranging from 0 ms to 1000 ms. Here we focus on an experimental condition in which retinal afterimages were suppressed by a phase shift towards the end of stimuli presentation. Validation of this method and results from the condition without a phase shift are provided in the Supplementary Information.

Experimental data. Recall error distributions and mean performance in Experiment 1 are plotted in Figs. 3A and B. Response error (measured as RMSE) increased with both set size and delay duration. A repeated measures ANOVA revealed a significant effect of set size ($F_{(2,18)} = 117.8, p < .001, \eta^2 = .44$), delay time ($F_{(5,45)} = 52, p < .001, \eta^2 = .23$), and their interaction ($F_{(10,90)} = 26.7, p < .001, \eta^2 = .13$) on response error. We further explored this interaction, first finding response error in the 1 item condition (red in Fig. 3) did not change with delay ($F_{(5,45)} = 1.32, p = .27, \eta^2 = .07$). This was supported by Bayesian analysis ($BF_{10} = 0.34$) which found weak to moderate evidence against modulation of 1 item recall by memory delay. In contrast, response error increased with delay for the remaining two set sizes (4 items, green; 10 items, blue; main effect: $F_{(5,45)} = 55, p < .001, \eta^2 = .48$). This increase in response error consisted of an initial rapid rise (over the first 200 ms), followed by a more gradual increase as the delay between stimulus and cue increased. Next, we found a modulating effect of delay on recall for the remaining two set sizes (interaction: $F_{(5,45)} = 10.1, p < .001, \eta^2 = .05$). The direct comparison revealed that the increase in response error with delay ($\Delta\text{RMSE} = \text{RMSE}_{1000\text{ms}} - \text{RMSE}_{\text{Simult}}$) was greater when observers memorized more items ($t_{(9)} = 9.1, p < .001, d = 2.88$).

One surprising result was the observed set size effect in the 0 ms delay condition ($F_{(2,18)} = 23.7, p < .001, \eta^2 = .53$) consistent with a stepwise increase in recall error with set size (pairwise comparison, $t_{(9)} \geq 2.88, p \leq .036, d \geq 0.91$, Bonferroni correction applied). Importantly, this effect was a consequence of responding based on a memory of the stimulus, since orientation reproduction was comparable across set sizes in the perceptual condition (simultaneous presentation; $F_{(2,18)} = 1.26, p = .3, \eta^2 = .04, BF_{10} = 0.47$). Previous studies have characterized iconic memory as an effectively unlimited store, capable of holding any number of items without a consequent loss of fidelity [58, 28]. While our modelling ultimately affirmed this conception of IM, we nonetheless show that recall of information is contingent on the number of objects concurrently in memory from the moment stimuli physically disappear (see below).

Taken together, these results provide evidence that the fidelity of stored representations changes dramatically over the first few moments after stimuli offset. We next aimed to explain the neural computations supporting these dynamics. In summary, behavioral data displayed three key characteristics we aimed to explain, all visible in Fig. 3B. First, recall fidelity for a single item remained relatively stable across changes in delay, and was the same as perceptual fidelity. Second, recall

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

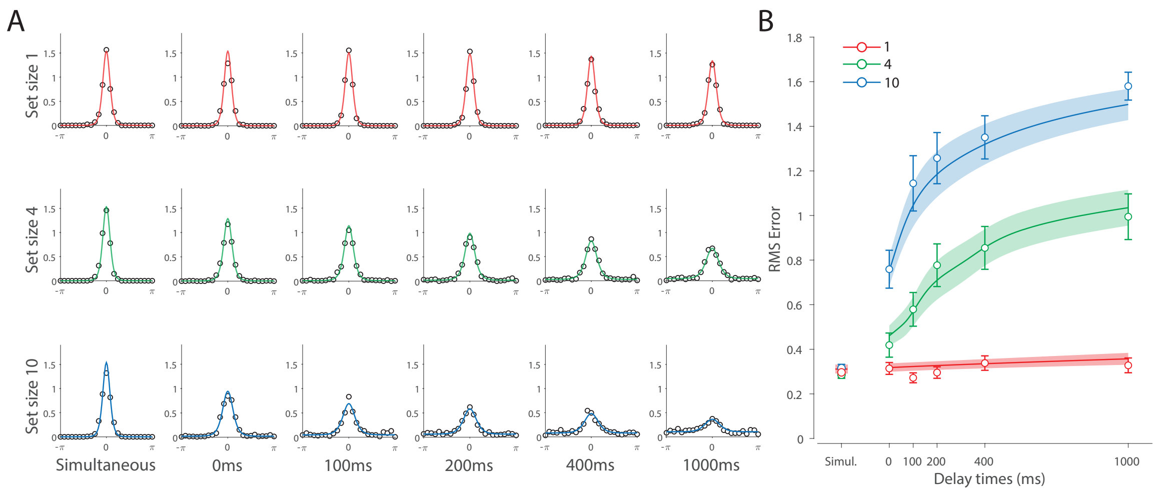

Figure 3: Experiment 1 data and model fits show the consequences of varying set size and delay duration on WM reproduction error. (**A**) Empirical recall error distributions (black circles) and the DyNR model fits (colored curves). Different panels correspond to different set sizes (rows) and delays (columns). (**B**) Corresponding RMS errors from experimental data (circles and errorbars) and the DyNR model fits (curves and error patches). Error bars and patches indicate ± 1 SEM.

fidelity for higher set sizes showed substantial, non-linear temporal dynamics. Lastly, recall fidelity was contingent on the number of stored items from the moment stimuli disappeared.

Dynamic Neural Resource model. Curves in Figs. 3A and B show fits of the model with maximum likelihood (ML) parameters (mean ± SE: population gain γ = 59.8 ± 3.3, tuning width κ = 3.21 ± 0.2, sensory decay time constant τ_{decay} = 0.21 ± 0.052, VWM accumulation time constant τ_{WM} = 0.096 ± 0.045, cue processing constant b = 0.171 s ± .055 s, base diffusion σ^2_{diff} = 0.03 ± 0.017, swap probability p = 0.027 ± 0.009). The model provided a close fit to response error distributions (Fig. 3A) and summary statistics (Fig. 3B; see also Fig. S3 for reproduction of swap error frequencies), successfully reproducing the pattern of changes with set size and delay. In particular, the model accounted for the three key observations identified above.

First, the model predicted the near-constant recall fidelity observed for a single item across these short retention intervals. The neural signal associated with the target object at recall depends on the normalized signal in VWM at offset supplemented by the available sensory signal post-cue. The sensory signal is integrated into VWM after the cue to fill any unallocated neural resource that arose by discarding uncued items. In the case of a single item, the entirety of VWM resources are allocated to one object during encoding, so no resource is freed by the cue that would allow the signal to be further strengthened based on the decaying sensory representation.

Importantly, this prediction contradicts the classical view of direct read-out from IM, according to which representational fidelity should be enhanced with very short delays irrespective of VWM limitations (see *Alternative accounts* below for a formal test of such a model). Note that the DyNR model nonetheless predicts some deterioration in fidelity over time even for a single item, due to noise-driven diffusion of the stored value. However, based on previous reports, we expected this process to be substantially slower and the impact on single item precision relatively small on this (≤1 s) timescale. The fitted diffusion parameters and resulting shallow slope of fitted RMS error (red curve in Fig. 3B) confirmed this.

Second, the neural model predicts the specific pattern of dynamics observed in trials with multiple items (set sizes 4, green, and 10, blue curves). Once the cue is presented, resources encoding uncued items are freed and the decaying sensory signal representing the target item is further integrated into VWM, still subject to limited total VWM resources but now without competition from other items. Due to exponential decay of the sensory signal, the increase in fidelity thus accrued changes rapidly with retention interval over the first few hundred milliseconds. At longer delays, the cue identifies

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the target only after the sensory signal has effectively disappeared, so the VWM signal representing the target item remains at the normalized level reflecting equal distribution between all items in the memory array, and memory dynamics consist only of the more gradual deterioration of fidelity due to accumulated noise in the encoded value.

Finally, the DyNR model predicts the presence of a set size effect on fidelity throughout the entire memory period, including the no delay (0 ms) condition in which the cue onset was coincident with stimulus offset. In the model, this behavior emerges as a consequence of two independent processes. First, at the end of stimulus presentation, items within smaller (lower set size) arrays are encoded in VWM with higher signal amplitude, reflecting normalization. This signal strength represents a baseline that can be supplemented by further integration of the sensory signal after an early cue. However, if the sensory decay is sufficiently rapid, then even if the cue is presented immediately the target representation will not attain the maximum amplitude (equivalent to set size of one) starting from a lower baseline. Second, as described by Hick's Law [53] it takes longer to identify the target item based on the cue as the number of alternatives increases (see *Alternative models* below for a formal test of this assumption). As a result, for higher set sizes, less sensory signal encoding the target item remains to be integrated into VWM once it has been identified.

Model variants. We next focused on alternative explanations for the temporal dynamics observed in Experiment 1. Specifically, we examined whether the observed dynamics could be accounted for either solely by post-stimulus changes in neural signal amplitude or solely by noise-driven diffusion of stored values. To pre-empt our conclusions, we demonstrate that both components are needed to explain the observed dynamics in memory fidelity. Moreover, to more closely examine the role of diffusion in WM dynamics, we fit our neural model to an additional dataset collected in our lab ([57]; see Additional dataset 1 in Supplementary Information). This experiment used longer delays compared to those used in Experiment 1, and therefore precluded any beneficial effect of post-stimulus sensory information, while at the same time allowing the diffusion to operate over a longer period. This experiment allowed us to test whether diffusion is sufficient to account for human recall errors with longer memory delays.

Fixed neural signal. A recent computational study on forgetting in VWM proposed that diffusion is sufficient to explain memory dynamics over delay [10]. To test for this, we developed two reduced versions of the DyNR model in which the diffusion process was solely responsible for memory fidelity dynamics. In both variants, the sensory signal terminated abruptly with stimuli offset, so the VWM signal encoding the stimuli was independent of the delay duration and equal to the limit imposed by normalization ($\check{\gamma}_{\mathrm{wm}}/N$). In the first variant, the diffusion rate was constant across set sizes, as in the full model. The formal model comparison demonstrated that the full DyNR model performed better than this simplified alternative (ΔAIC $= 609.5$).

In the second variant, we allowed the diffusion rate to increase proportionally with set size (for a similar proposal see [59]). This model was again outperformed by the full DyNR model (ΔAIC $= 666.4$). Critically, both models tested here failed to qualitatively reproduce the observed non-linear pattern of changes in recall error with time, notably overestimating recall error at the shortest delays by assuming no modulation in the representational signal (Fig. S4).

Diffusion. We developed two variants of the proposed neural model to test the role of diffusion. In the first variant, we completely omitted the diffusion process from the model to test whether the sensory signal modulation during the retention period is sufficient to explain temporal dynamics in recall fidelity. It could be argued that diffusion accounts for only minor changes in precision over brief delays as used here, and therefore adds unnecessary complexity to the proposed model without improving the fit substantially. However, the formal model comparison revealed that the full DyNR model provides a better fit to human recall error compared to the matching model without diffusion (ΔAIC $= 17.9$).

The second variant was identical to the proposed model, except that we replaced the constant diffusion rate with a set size scaled diffusion rate (see Eq 10). The model comparison showed that the full DyNR model also outperformed this variant (ΔAIC $= 29.8$). While both model variants qualitatively reproduced the increase in memory error with delay and set size, the pattern of variability

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

was better explained by the model with a constant diffusion rate across set sizes. Although a more substantial diffusion effect could become apparent with longer delays than those used here, previous work demonstrated that noise-driven diffusion causes representations to deteriorate throughout the entire retention period [55].

Finally, we examined the role of diffusion with longer memory intervals in a separate experiment using variable set sizes and memory intervals (1 and 7 seconds) (for full details see Additional dataset 1 in Supplementary Information). We demonstrated that, once sensory information decayed completely, an accumulation of error during retention interval accounted for continuing memory deterioration. Together, the results presented here corroborate findings on the role of diffusion in temporal dynamics of recall fidelity [9].

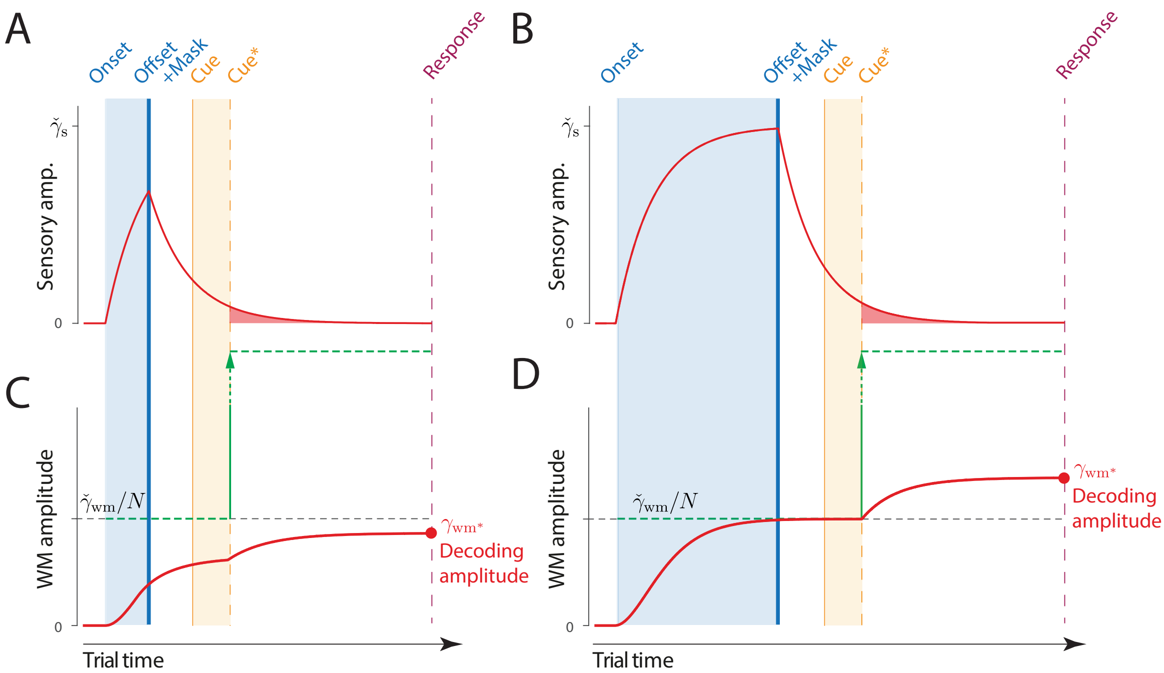

Figure 4: Time course of sensory and WM gain with variable exposure duration.
(**A, B**) The signal amplitude in the sensory population increases from stimulus onset, exponentially approaching the maximum sensory activity ($\tilde{\gamma}_s$). For shorter presentation durations (A) the attained amplitude at stimulus offset is only a fraction of the maximum (compare B, late offset). Following offset, sensory areas produce a decaying neural response, that is curtailed (faster decay) but not abolished by a backward mask.
(**C, D**) Information about the stimulus is accumulated in WM from sensory activity. A shorter presentation (C) provides less sensory evidence for the initial accumulation of all items into VWM (compare **D**, late offset), and subsequently less decaying sensory activity that can supplement VWM activity for the target item following the cue.

Experiment 2: Exposure duration

In Experiment 2, we evaluated the encoding phase of VWM, by testing recall of orientation stimuli displayed in arrays of variable size presented for variable durations. In the DyNR model, increasing the sensory evidence by prolonging stimulus presentation has a favorable effect on later recall of stimulus, as more of that evidence can be accumulated into VWM. Importantly, this accumulation is also capped by the VWM resources available to store it.

Experimental data. Figure 5 shows the response error for different presentation durations and set sizes. Consistent with previous findings, response error can be seen to decrease with prolonged presentation duration, but increase as the number of items in memory increases. This was confirmed with a significant effect of display duration ($F_{(6,72)} = 29.01, p < .001, \eta^2 = .21$), set size ($F_{(2,24)} = 112.51, p < .001, \eta^2 = .54$), and their interaction ($F_{(12,144)} = 2.58, p = .004, \eta^2 = .019$).

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We further explored this interaction by first confirming that response error decreased with display duration within each set size ($F_{(6,72)} \geq 10.24, p < .001, \eta^2 \geq .26$). A consistent pattern was observed across set sizes, comprising an initial rapid decrease in response error over the briefest presentation times (first 200 ms), followed by saturation at prolonged exposure durations. Next, we calculated the change in recall error between the longest and the shortest display exposure within each set size, revealing that response error decreased more rapidly with display time as the number of items in memory decreased (ANOVA: $F_{(2,24)} = 7.79, p = .002, \eta^2 = .21$; corrected pairwise comparisons: $t_{1-4} = 3.65, p = .016, d = 0.87$, $t_{4-10} = 0.96, p = .72, d = .27$).

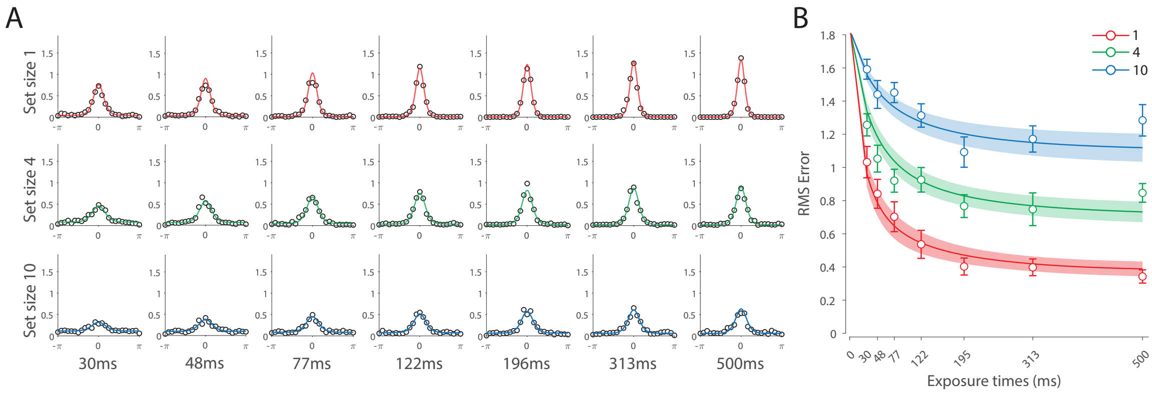

Figure 5: Experiment 2 results and modelling data show the consequences of varying set size and stimulus exposure time on VWM reproduction error. (**A**) Empirical recall error distributions (black circles) and the DyNR model fits (colored curves). Different panels correspond to different set sizes (rows) and exposure durations (columns). (**B**) Corresponding RMS errors from experimental data (circles and errorbars) and the DyNR model fits (curves and error patches). Error bars and patches indicate ± 1 SEM.

These results reveal the time course of information accumulation into VWM and forming of stable representations. We again identified several key characteristics of the dynamics of recall fidelity in the data (Fig. 5B) to test agaist the DyNR model. Consistent with previous studies, we found recall fidelity changed with both presentation duration and the number of presented stimuli [35–37]. Specifically, as display duration increased from the shortest exposure, recall error showed an initial rapid decrease followed by a gradual levelling-off. As set size increased, the initial slope became shallower and the plateau occurred at a higher level of error.

Dynamic Neural Resource model. Curves in Figs. 5A & B show fits of the model with maximum likelihood (ML) parameters (mean ± SE: population gain $\gamma = 188.5 \pm 109.6$, tuning width $\kappa = 10.2 \pm 6.08$, sensory rise time constant $\tau_{\text{rise}} = 0.33 \pm 0.18$, sensory decay time constant $\tau_{\text{decay}} = 0.61 \pm 0.19$, VWM accumulation time constant $\tau_{\text{WM}} = 0.8 \pm 0.34$, cue processing constant $b = 0.2$ s ± .09 s, base diffusion $\sigma^2_{\text{diff}} = 0.28 \pm 0.08$, spatial uncertainty time constant $\tau_{\text{spatial}} = 0.013 \pm 0.004$, swap probability $p = 0.053 \pm 0.01$). The model provided an excellent quantitative fit to response distributions (Fig. 5A) and RMSE (Fig. 5B), successfully reproducing the pattern of changes with set size and presentation duration.

The model predicted that information from a visible stimulus accrues at a high rate immediately after the stimulus onset, consistent with observed changes in human recall error over stimulus durations up to 200 ms (Fig. 5). This initial high encoding rate emerges naturally in the model due to the joint dynamics of sensory and VWM populations. In the sensory population, a low-pass temporal filter serves as a neural gain control mechanism, attenuating neural response to transient changes in stimuli [40, 43]. As a consequence, the neural response to stimulus onset increases exponentially (Fig. 4). The information from sensory areas is accumulated into VWM, such that the accumulation rate is directly proportional to the difference between the current and saturating state (i.e. the rate is faster when accumulated information is far from the saturating state). Therefore, dynamics in the sensory and VWM population jointly account for the initial high rate of information extraction from stimuli, and its dependence on set size.

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

After the initial steep change, the model predicts that recall fidelity will asymptote. This was again observed in human behavior (Fig. 5). Extending stimulus presentation beyond 200 ms had negligible impact on recall precision, consistent with previous studies [35]. The model explains this behavior by describing how sensory signal and VWM accumulation independently saturate with time. Since the temporal filtering in the sensory population attenuates only high-frequency stimuli (i.e. very short presentations), with sufficient exposure, the sensory signal plateaus, resulting in a stable feed-forward input to VWM. Similarly, VWM signal strength is subject to limits determined by normalization. Once the accumulated information reaches the normalized maximum set by the number of objects in memory, further accumulation of sensory evidence is not possible. Following the cue, a portion of the resource is freed, allowing the target representation to be further strengthened. However, because the sensory signal plateaus at longer exposures, the information available for integration after the cue remains constant across the longer exposures, supplementing normalized VWM signal by the same amount. The result is a plateau in fidelity that varies with set size.

Model variants. We investigated whether post-stimulus sensory persistence contributed to the model fits in Experiment 2. We assumed that the signal persisting after stimulus offset would be impaired but not eliminated by the subsequent presentation of a noise mask in this experiment [47]. An alternative account suggests that the mask immediately terminates any stimulus-related signal. To test for this, we fit a variant of the DyNR model in which the sensory signal was terminated by the onset of the mask, providing a feed-forward signal to VWM only for the period of the stimulus presentation. We found that the proposed DyNR model, in which some sensory signal persists after the mask onset, gave a better account of the data than this model variant (ΔAIC = 446.67). Although the alternative model captured the general pattern of changes in memory fidelity with exposure duration, it mispredicted fidelity at shorter exposures, in particular the effect of set size (Fig. S5A).

A testable prediction of this alternative model is that the memory fidelity at recall should obey the neural normalization principle because there was no additional signal to supplement the presentation after initial encoding. To test for this, we additionally fitted each exposure condition separately using the original Neural Resource model with only three parameters (i.e., neural gain, tuning width, and swap probability). This model failed to predict actual fidelity levels at recall (Fig. S5B), corroborating the findings of the model comparison.

Finally, to investigate the role of the post-stimulus sensory persistence on encoding dynamics, we fit the DyNR model to an additional dataset from Bays et al. [35] (for full details see Additional dataset 2 in Supplementary Information). This experiment aimed to investigate VWM dynamics during encoding, like our Experiment 2. In contrast to our Experiment 2, Bays et al. [35] used a much longer delay interval (1100 ms vs 100 ms), precluding the possibility of further accumulation of sensory evidence following the cue. We expected that the DyNR model could account for memory dynamics in this study without any post-stimulus sensory activity. This was confirmed by accurately reproducing memory dynamics with a model in which encoding into VWM relied only on sensory evidence during stimulus presentation (detailed results in Supplementary Information).

Alternative accounts

Having demonstrated the need for both post-stimulus sensory persistence and diffusion to account for empirical data, we next considered alternatives to our account of VWM accumulation and information read-out.

Direct read-out of sensory information. In the DyNR model, recall fidelity is enhanced following the cue by integrating remaining sensory activity into capacity-limited VWM. As a consequence, response precision is bounded from above by the memory limit irrespective of the available sensory signal. An alternative possibility is that the decaying sensory representation can be directly read out following the cue to inform a response, bypassing working memory limitations. To formalize this alternative model, we assumed that independent sensory and VWM representations would be

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

optimally combined via summation of neural activity to yield population gain

$$\gamma^*_{\text{sum}} = \gamma_{\text{wm}}(t_{\text{cue}^*}) + \gamma_{\text{s}}(t_{\text{cue}^*}) \tag{14}$$

The model is otherwise identical to the proposed DyNR model. A distinctive prediction of this model is that the precision of recall changes exponentially with delay for every set size, including 1 item (Fig. S8). This prediction is qualitatively inconsistent with the pattern of results observed in Experiment 1, in contrast with the DyNR model which does not predict any beneficial effect of earlier cues with set size 1. This alternative model provided a worse fit to data from Experiment 1 (ΔAIC = 164) and Experiment 2 (ΔAIC = 84.6), for combined evidence favouring the DyNR model of ΔAIC = 248.6.

Cue processing. In the DyNR model, we assumed that identifying the target stimulus based on the cue is time-consuming, and becomes more so as the number of alternatives increases. Cue processing time encompasses perceptual, attentional, and decision components needed to interpret and act on the cue. We tested the necessity of this component by fitting a model variant in which VWM started accumulating evidence about the cued item at the moment of cue presentation. This model provided a worse fit to empirical data from both Experiment 1 (ΔAIC = 84.5) and Experiment 2 (ΔAIC = 107.5), for total evidence in favor of the DyNR model of ΔAIC = 192 (Fig. S6). We fit another variant in which cue processing time was constant across set sizes. This alternative provided a worse fit to the data in Experiment 1 (ΔAIC = 191.6) and Experiment 2 (ΔAIC = 105), for combined evidence ΔAIC = 296.6 in favor of the full DyNR model that assumes cue processing time increases with set size. These results corroborate previous findings on the important role of cue processing time in models of attention [54] and IM [60].

Constant accumulation rate. In the DyNR model, the rate of accumulation into VWM is proportional to the difference between the present VWM amplitude and the maximum normalized amplitude (Eq. 2). An arguably simpler assumption is that the neural signal approaches saturation at a constant rate [61, 62]. In particular, the rate at which the signal representing an item is transferred to VWM is constant and depends only on the number of encoded items, i.e.

$$\dot{\gamma}_{\text{wm}}(t) = \begin{cases} \gamma_{\text{s}}(t)/(M(t)\tau_{\text{wm}}) & \text{if} \quad \gamma_{\text{wm}}(t) < \check{\gamma}_{\text{wm}}/M(t) \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

The dependence on $M(t)$ satisfies the constraint that the neural resources in VWM are allocated at a constant rate, irrespective of the number of items. We applied this model to psychophysical data from both experiments (Fig. S7) and found it provides a worse fit to the data from Experiment 1 (ΔAIC = 11.5) and Experiment 2 (ΔAIC = 36.2), for combined evidence favouring the DyNR model with exponential saturation (ΔAIC = 47.7).

Discussion

In the present study, we investigated the temporal dynamics of short-term recall fidelity. We conducted two new human psychophysical experiments and analyzed two existing datasets in order to characterize how recall errors are influenced by set size, stimulus duration and retention interval. We developed a Dynamic Neural Resource (DyNR) model to provide a mechanistic explanation of the observed behavior, capturing not only changes in overall fidelity but also the distribution of errors in the stimulus space and frequencies of swaps (intrusion errors). A key finding is that the benefit to recall precision observed at very short delays is due to additional post-cue integration of sensory information into working memory, and that direct retrieval from sensory memory is unable to account for the empirical patterns of error.

Sensory and WM dynamics during delay

In the first experiment we investigated the effects of brief unfilled delays on recall fidelity. With multi-item arrays, we observed that memory performance deteriorates precipitously over the first few hundred milliseconds after stimuli disappear, followed by a gradual levelling-off of error with longer delays (Fig. 3). These results are consistent with previously reported patterns of memory

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

dynamics [27–29, 31], and estimates of sensory decay ranging between 100 ms and 400 ms [63, 64]. Here, we shed new light on these results by taking a computational approach in explaining observed temporal dynamics, and asking what this superior recall's neural origin is and its relation with VWM. To answer these questions, we adapted the Neural Resource model of Bays [12] with a temporal component. The new DyNR model considers dynamics in a sensory neural population registering the stimuli and in a VWM population that stores the stimuli for later recall. Critically, our model assumes that objects encoded with limited precision into VWM can be flexibly supplemented with sensory activity following a recall cue, within a brief temporal window while the sensory population provides a feed-forward input post-stimulus. The boost in the representational VWM signal predicts a behavioral benefit of early cues that is consistent with our data and a large corpus of previous experiments [51].

A common assumption in studies of visual short-term memory is that recall over brief delays is exclusively supported by one of two memory stores, IM or VWM [29, 30]. In this account, a cue presented within the first few hundred milliseconds after stimulus offset allows observers to access high resolution but rapidly deteriorating representations in IM; once the information in IM has decayed, objects must be retrieved from the capacity-limited VWM store. Two pieces of evidence from the current study contradict this view and strongly suggest that recall depends on VWM from the moment objects disappear. First, the recall benefit of short delays was not observed for one item arrays. We propose that this behavior reflects the fact that, during encoding, the entirety of the VWM resource is allocated to a single object, leaving no free capacity for further enhancement based on the available sensory signal post-cue. Second, we found clear evidence that recall fidelity varied with set size even with no delay between stimulus offset and cue (0 ms condition). We argue that this arises from the set-size dependence of representational fidelity in VWM, which is only incompletely compensated by integration of the decaying sensory signal post-cue, resulting in lower fidelity for higher set sizes. The DyNR model provides a successful quantitative account for these findings, which are in clear contrast with the traditional view of IM.

The rapid changes in fidelity over short delays can be distinguished from dynamics over longer retention intervals. A number of recent studies have observed a slow deterioration of VWM precision over the course of prolonged retention [9, 21, 22, 65–67]. The causes of this deterioration are still contested, but growing evidence links this behavior to noise-driven diffusion. At a mechanistic level, diffusion is considered a fundamental property of continuous attractor networks of the kind commonly associated with models of working memory [68, 69]. In such networks, memorized features are represented as persistent activity "bumps" in the network's representational feature space. Over a memory delay, the activity bump is sustained by balanced excitatory and inhibitory connections, while stochasticity in neural activity causes shifts of the bump along the feature dimension, taking the form of a random walk. Although we did not model the network processes governing stability and diffusion within neural populations, our implementation is consistent with random (Brownian) perturbation, as assumed by attractor models (see also 9).

Our theoretical account of memory dynamics during delay differs from several existing models of forgetting, which emphasize diffusion as the dominant source of error in short-term memory (e.g., 10, 59). To solely account for the observed data in Experiment 1, diffusion would need to be strongest early in the retention period, followed by a much weaker diffusion with longer delays. However, it is unclear why the diffusion rate would change, and particularly slow down, with time. Assuming a constant neural signal encoding the stimulus, this would predict greater variability in neural activity initially compared to the later period after stimuli offset. This is inconsistent with electrophysiological data showing relatively stable levels of spiking variability throughout the memory delay period [70, 71]. The results observed here are consistent with the proposal that modulation of neural signal over short memory intervals accounts for an abrupt change in response fidelity, while diffusion accounts for a slower change that grows with time.

In the present study, a model assuming a constant diffusion rate, independent of the stored number of items, was preferred to one in which diffusion rate increases linearly with set size. This is consistent with results of Shin et al. [66] who did not find a significant effect of set size on the rate of memory deterioration. In contrast to that, Koyluoglu et al. [59] recently proposed that the rate of diffusion scales with set size. However, this study did not account for the presence of swap errors, which we

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

found to increase with retention interval as well as set size. To draw strong conclusions about the dependence of diffusion on set size would require a future study to disentangle the different sources of error that could, in principle, increase with delay.

Sensory and WM dynamics during encoding

Having investigated memory degradation during the retention interval, in Experiment 2 we focused on the dynamics arising from accumulation of information during stimulus presentation. Using new psychophysical data, we showed that encoding of information into VWM is contingent on both presentation duration and the number of memorized stimuli. The observed patterns of data indicate that VWM encoding of elementary stimuli is mostly completed within the first 200 ms of presentation even at the largest set sizes, with minimal benefit of longer exposures, extending previous work [35–37]. This fast encoding process may have an adaptive role: with a key function of VWM to store and accumulate information across saccadic eye movements, an efficient system should deploy its resources within the duration of a typical gaze fixation [72].

Our aim was again to move beyond the description of the encoding dynamics and to provide a biologically plausible neurocomputational account of these dynamics. To achieve that, we applied the same VWM accumulation process that operates post-cue to the sensory information during stimulus presentation. Using previously published and newly collected data, we show that a model in which VWM accumulates dynamical sensory input up to a fidelity limit can successfully account for patterns of human recall errors with variable set size and stimulus presentation. An important result of our modelling is that the accumulated information in VWM increases with a rate proportional to unfilled capacity. In particular, the model with such exponential accumulation provided a better fit than a model assuming a constant encoding rate. This parallels previous observations that models based on exponential-like extraction of information successfully characterize attention [73, 74], working memory encoding [35, 75], memory updating [76], and broader cognitive processes [77]. We hypothesize that this pattern represents an approach to an equilibrium state of balanced excitation from the sensory input and lateral inhibition within the VWM population, which is the basis for capacity of the memory system.

Our computational account of VWM encoding dynamics differs from several existing modelling frameworks aiming to explain similar data. For example, the Theory of Visual Attention (TVA; 73) assumes that visual stimuli participate in a parallel exponential race towards limited VWM. Like the DyNR model, TVA assumes a form of normalization in the sense that the speed with which items race towards VWM depends on the number of items in the visual field. Unlike our dynamic model, TVA is not a theory of VWM, and it considers VWM only as a storage for categorizations of visual objects. In particular, TVA takes into account the limits of VWM but does not specify why or how these limitations arise. Finally, TVA considers whether an object was selected for entry into VWM in an all-or-none fashion; our dynamic model is mostly concerned with the fidelity of representations. A somewhat alternative account of VWM encoding is provided by the Competitive Interaction Theory (CIT; 78), which is similarly based on the Signal Detection theory and principles of normalization [39]. Like TVA, CIT is mostly focused on item selection and merely incorporates a concept of VWM capacity derived from object-based models of VWM. Although CIT had success in accounting for behavioral data from a two-alternative orthogonal discrimination task using up to four items and a limited range of encoding times, it remains an open question whether this model can account for error distributions as measured in a continuous report task, and a larger range of set sizes and stimulus exposures. Importantly, compared to both TVA and CIT, the DyNR model is strongly rooted in and inspired by findings from neuroscience. This not only adds to the biological plausibility of our model but also allows future studies to test the model's predictions using physiological methods.

Neural mechanisms

The theory presented here generalizes the Neural Resource model of Bays [12], a simple encoding-decoding model in which visual features are represented in the noisy spiking activity of neural populations [15], and where the activity representing each feature scales inversely with the total number of representations, consistent with the prevalence of normalization mechanisms in the brain and observations from single-neuron recording [79] and fMRI decoding [80] studies. The

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

population coding in the model is based on an abstract idealization of neural response functions. Nevertheless, it has recently been shown that more realistic population coding schemes that allow for heterogeneity in neural tuning curves and correlated spiking activity as observed in visual cortex, maintain the key predictions of the idealized model [81, 13]. This may be seen as a consequence of the different population codes inducing a common representational geometry [82].

We adapted the stationary VWM model by first incorporating a sensory population that provides an input drive to the VWM population. In parallel with neurophysiological observations, a common approach is to model these dynamics with a low-pass filter which acts like a neural gain modulation mechanism [43]. As a consequence, the sensory response to stimulus onset and offset is an exponential rise and decay in activity, respectively. The decaying component of the response has been recognized as a neural substrate of visual persistence and IM [34, 33]. Here, we modelled sensory decay with an exponential function [83], although other forms of decay have been proposed. For example, Loftus et al. [63] showed that iconic decay could be better captured using a gamma survival function, a generalization of exponential decay that could simply be implemented in our neural model by replacing a single filter with a cascade of exponential low-pass filters.

In addition to the dynamics in the sensory population, two features of VWM introduce additional dynamics in representation fidelity: the accumulation of information (discussed above) and the diffusion of representations owing to accumulated noise. Although we did not aim to model the neural processes behind diffusion, our implementation is consistent with the consequences of neural variability in attractor networks [23, 69]. Converging neural evidence demonstrating such diffusion has been observed using single-unit neural recording in monkeys [24], as well as EEG [26] and fMRI [25, 84] studies in humans.

Our model makes a clear distinction between dynamics in sensory and VWM populations, however, it remains agnostic as to whether the populations have the same or different anatomical locus [85]. Albeit inspired by the properties of orientation-selective neurons in area V1, population tuning of this kind is a common coding motif across the brain [15]. While it could be considered efficient to use already specialized circuits to maintain as well as process visual information, it is still debated whether sensory areas are a feasible candidate for memory storage [86, 87]. While some studies have focused on prefrontal [88], parietal [89] or occipital [90] cortices as the primary locus of VWM, others argue for distributed storage by demonstrating that VWM contents can be decoded from imaging signals originating in multiple brain areas [91].

Representational dynamics of cue-dimension features

Memory retrieval failures in which a non-cued item is reported in place of the intended target represent an important source of error in VWM recall. These swap errors occur more often at higher set sizes and when spatial confusability is high [92, 93], as predicted by models in which they arise from uncertainty in the recall of cue-dimension features leading to incorrect selection of an item in memory [19, 56]. In the current study, we assumed memory for spatial location (the cue feature) undergoes similar dynamics to memory for orientation (the report feature), and in particular that spatial information degrades with retention time [9], leading to changes in swap error frequency with delay interval. Similarly, during encoding the fidelity of spatial representation increases with the accumulation of sensory evidence [94], reducing the uncertainty at retrieval and consequently swap errors at longer stimulus exposure. Although we did not explicitly model the neural signals representing location, the modelled dynamics in the probability of swap errors were consistent with those of the primary memory feature. Future studies might develop and test more detailed models of the cue identification process based on how swap frequency changes with time.

Removal of information from WM

In the DyNR model, taking advantage of early cues requires rapid removal of the VWM signal associated with uncued items, to admit further accumulation of activity encoding the cued item. To achieve this, an active process of selective content elimination may be required [52], as opposed to a passive decay of uncued representations during the post-cue interval. Mounting evidence for such active removal has been provided at the behavioral [95] and neural [96] level. Importantly, studies

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

show that a functional role of such active removal is to release resources allocated to the uncued representations, facilitating the encoding of new information [97]. The fast reallocation of neural resources assumed by the DyNR model is consistent with such a description of active removal.

Data Availability

Data and code related to this study will be made available at https://doi.org/10.17863/CAM.95223.

Acknowledgments

We thank George Sperling and Sebastian Schneegans for helpful discussion, and Robert Taylor for help with Bayesian hierarchical modelling. This research was supported by the Wellcome Trust (grant 106926 to PMB).

References

[1] Mark D'Esposito and Bradley R. Postle. The Cognitive Neuroscience of Working Memory. *Annual Review of Psychology*, 66(1):115–142, 2015. doi: 10.1146/annurev-psych-010814-015031.

[2] Tatiana Pasternak and Mark W. Greenlee. Working memory in primate sensory systems. *Nature Reviews Neuroscience*, 6(2):97–107, 2005. ISSN 1471-003X. doi: 10.1038/nrn1603.

[3] Wei Ji Ma, Masud Husain, and Paul M Bays. Changing concepts of working memory. *Nature Neuroscience*, 17(3):347–356, 2014. ISSN 1097-6256, 1546-1726. doi: 10.1038/nn.3655.

[4] Paul M. Bays, Sebastian Schneegans, Wei Ji Ma, and Timothy Brady. Representation and computation in working memory. October 2022. doi: 10.31234/osf.io/kubr9.

[5] H. B. Barlow. The Ferrier Lecture, 1980. Critical limiting factors in the design of the eye and visual cortex. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 212 (1186):1–34, 1981. ISSN 0950-1193. doi: 10.1098/rspb.1981.0022.

[6] D. C. Van Essen, C. H. Anderson, and D. J. Felleman. Information processing in the primate visual system: an integrated systems perspective. *Science (New York, N.Y.)*, 255(5043):419–423, 1992. ISSN 0036-8075.

[7] Bingni W. Brunton, Matthew M. Botvinick, and Carlos D. Brody. Rats and Humans Can Optimally Accumulate Evidence for Decision-Making. *Science*, 340(6128):95–98, 2013. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1233912.

[8] Joshua I. Gold and Michael N. Shadlen. The Neural Basis of Decision Making. *Annual Review of Neuroscience*, 30(1):535–574, 2007. ISSN 0147-006X, 1545-4126. doi: 10.1146/annurev.neuro.29.051605.113038.

[9] Sebastian Schneegans and Paul M. Bays. Drift in Neural Population Activity Causes Working Memory to Deteriorate Over Time. *The Journal of Neuroscience*, 38(21):4859–4869, 2018. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.3440-17.2018.

[10] Matthew F. Panichello, Brian DePasquale, Jonathan W. Pillow, and Timothy J. Buschman. Error-correcting dynamics in visual working memory. *Nature Communications*, 10(1), 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11298-3.

[11] Freek van Ede, Sammi R. Chekroud, Mark G. Stokes, and Anna C. Nobre. Concurrent visual and motor selection during visual working memory guided action. *Nature Neuroscience*, 2019. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-018-0335-6.

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

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[13] 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*, page 202004306, 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2004306117.

[14] 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. ISSN 1097-6256, 1546-1726. doi: 10.1038/nn1790.

[15] A. Pouget, P. Dayan, and R. Zemel. Information processing with population codes. *Nature Reviews. Neuroscience*, 1(2):125–132, 2000. ISSN 1471-003X. doi: 10.1038/35039062.

[16] M Carandini and D.J. Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1), 2011. ISSN 1471-003X, 1471-0048. doi: 10.1038/nrn3136.

[17] Ivan Tomić and Paul M. Bays. Perceptual similarity judgments do not predict the distribution of errors in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2022. ISSN 1939-1285, 0278-7393. doi: 10.1037/xlm0001172.

[18] Paul M. Bays and Robert Taylor. A neural model of retrospective attention in visual working memory. *Cognitive Psychology*, 100:43–52, 2018. ISSN 00100285. doi: 10.1016/j.cogpsych.2017.12.001.

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

[20] Paul M. Bays. A signature of neural coding at human perceptual limits. *Journal of Vision*, 16 (11):4, 2016. ISSN 1534-7362. doi: 10.1167/16.11.4.

[21] Yoni Pertzov, Sanjay Manohar, and Masud Husain. Rapid forgetting results from competition over time between items in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(4):528–536, 2017. ISSN 1939-1285, 0278-7393. doi: 10.1037/xlm0000328.

[22] Rosanne L. Rademaker, Young Eun Park, Alexander T. Sack, and Frank Tong. Evidence of gradual loss of precision for simple features and complex objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, 2018. ISSN 1939-1277, 0096-1523. doi: 10.1037/xhp0000491.

[23] Y. Burak and I. R. Fiete. Fundamental limits on persistent activity in networks of noisy neurons. *Proceedings of the National Academy of Sciences*, 109(43):17645–17650, 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1117386109.

[24] 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, 2014. ISSN 1097-6256, 1546-1726. doi: 10.1038/nn.3645.

[25] Phui Cheng Lim, Emily J. Ward, Timothy J. Vickery, and Matthew R. Johnson. Not-so-working Memory: Drift in Functional Magnetic Resonance Imaging Pattern Representations during Maintenance Predicts Errors in a Visual Working Memory Task. *Journal of Cognitive Neuroscience*, pages 1–15, 2019.

[26] Michael J. Wolff, Janina Jochim, Elkan G. Akyürek, Timothy J. Buschman, and Mark G. Stokes. Drifting codes within a stable coding scheme for working memory. *PLOS Biology*, 18 (3):e3000625, 2020. ISSN 1545-7885. doi: 10.1371/journal.pbio.3000625.

[27] Vincent Di Lollo and Peter Dixon. Two forms of persistence in visual information processing. *Journal of Experimental Psychology: Human Perception and Performance*, 14(4):671–681, 1988. ISSN 1939-1277, 0096-1523. doi: 10.1037/0096-1523.14.4.671.

[28] George Sperling. The information available in brief visual presentations. *Psychological Monographs: General and Applied*, 74(11):1–29, 1960. ISSN 0096-9753. doi: 10.1037/h0093759.

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[29] Claire Bradley and Joel Pearson. The Sensory Components of High-Capacity Iconic Memory and Visual Working Memory. *Frontiers in Psychology*, 3, 2012. ISSN 1664-1078. doi: 10.3389/fpsyg.2012.00355.

[30] Michael S. Pratte. Iconic Memories Die a Sudden Death. *Psychological Science*, 29(6):877–887, 2018. ISSN 0956-7976, 1467-9280. doi: 10.1177/0956797617747118.

[31] Ulric Neisser. *Cognitive Psychology*. Number 1966 in Century psychology series award. Appleton-Century-Crofts, New York, NY, 1967. ISBN 978-0-390-66509-6.

[32] Edmund T. Rolls and M.J. Tovee. Processing speed in the cerebral cortex and the neurophysiology of visual masking. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 257(1348):9–15, 1994. ISSN 0962-8452, 1471-2954. doi: 10.1098/rspb.1994.0087.

[33] Rob R.M. Teeuwen, Catherine Wacongne, Ulf H. Schnabel, Matthew W. Self, and Pieter R. Roelfsema. A neuronal basis of iconic memory in macaque primary visual cortex. *Current Biology*, 31(24), 2021. ISSN 09609822. doi: 10.1016/j.cub.2021.09.052.

[34] Timo van Kerkoerle, Matthew W. Self, and Pieter R. Roelfsema. Layer-specificity in the effects of attention and working memory on activity in primary visual cortex. *Nature Communications*, 8(1), 2017. ISSN 2041-1723. doi: 10.1038/ncomms13804.

[35] P. M. Bays, N. Gorgoraptis, N. Wee, L. Marshall, and M. Husain. Temporal dynamics of encoding, storage, and reallocation of visual working memory. *Journal of Vision*, 11(10):6–6, 2011. ISSN 1534-7362. doi: 10.1167/11.10.6.

[36] Hitomi Shibuya and Claus Bundesen. Visual selection from multielement displays: Measuring and modeling effects of exposure duration. *Journal of Experimental Psychology: Human Perception and Performance*, 14(4):591–600, 1988. ISSN 1939-1277, 0096-1523. doi: 10.1037/0096-1523.14.4.591.

[37] Edward K. Vogel, Geoffrey F. Woodman, and Steven J. Luck. The time course of consolidation in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, 32(6):1436–1451, 2006. ISSN 1939-1277, 0096-1523. doi: 10.1037/0096-1523.32.6.1436.

[38] Mehrdad Jazayeri and J Anthony Movshon. Optimal representation of sensory information by neural populations. *Nature Neuroscience*, 9(5):690–696, 2006. ISSN 1097-6256. doi: 10.1038/nn1691.

[39] John H. Reynolds and David J. Heeger. The Normalization Model of Attention. *Neuron*, 61 (2):168–185, 2009. ISSN 08966273. doi: 10.1016/j.neuron.2009.01.002.

[40] R. F. Hess and R. J. Snowden. Temporal properties of human visual filters: number, shapes and spatial covariation. *Vision Research*, 32(1):47–59, 1992. ISSN 0042-6989. doi: 10.1016/0042-6989(92)90112-v.

[41] A M Derrington and P Lennie. Spatial and temporal contrast sensitivities of neurones in lateral geniculate nucleus of macaque. *The Journal of Physiology*, 357(1):219–240, 1984. ISSN 00223751. doi: 10.1113/jphysiol.1984.sp015498.

[42] B B Lee, P R Martin, and A Valberg. Sensitivity of macaque retinal ganglion cells to chromatic and luminance flicker. *The Journal of Physiology*, 414(1):223–243, 1989. ISSN 00223751. doi: 10.1113/jphysiol.1989.sp017685.

[43] M. J. Hawken, R. M. Shapley, and D. H. Grosof. Temporal-frequency selectivity in monkey visual cortex. *Visual Neuroscience*, 13(3):477–492, 1996. ISSN 0952-5238. doi: 10.1017/s0952523800008154.

[44] J. R. Müller, A. B. Metha, J. Krauskopf, and P. Lennie. Information conveyed by onset transients in responses of striate cortical neurons. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 21(17):6978–6990, 2001. ISSN 1529-2401.

[45] M. W. Oram and D. I. Perrett. Time course of neural responses discriminating different views of the face and head. *Journal of Neurophysiology*, 68(1):70–84, 1992. ISSN 0022-3077, 1522-1598. doi: 10.1152/jn.1992.68.1.70.

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[46] Yiu Fai Sit, Yuzhi Chen, Wilson S. Geisler, Risto Miikkulainen, and Eyal Seidemann. Complex Dynamics of V1 Population Responses Explained by a Simple Gain-Control Model. *Neuron*, 64 (6):943–956, 2009. ISSN 08966273. doi: 10.1016/j.neuron.2009.08.041.

[47] G. Kovacs, R. Vogels, and G. A. Orban. Cortical correlate of pattern backward masking. *Proceedings of the National Academy of Sciences*, 92(12):5587–5591, 1995. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.92.12.5587.

[48] M. T. Turvey. On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with patterned stimuli. *Psychological Review*, 80(1):1–52, 1973. ISSN 1939-1471, 0033-295X. doi: 10.1037/h0033872.

[49] E. Aksay, G. Gamkrelidze, H. S. Seung, R. Baker, and D. W. Tank. In vivo intracellular recording and perturbation of persistent activity in a neural integrator. *Nature Neuroscience*, 4(2):184–193, 2001. ISSN 1097-6256. doi: 10.1038/84023.

[50] A. Compte, N. Brunel, P. S. Goldman-Rakic, and X. J. Wang. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex (New York, N.Y.: 1991)*, 10(9):910–923, 2000. ISSN 1047-3211.

[51] Max Coltheart. Iconic memory and visible persistence. *Perception & Psychophysics*, 27(3): 183–228, 1980. ISSN 0031-5117, 1532-5962. doi: 10.3758/BF03204258.

[52] Klaus Oberauer. Removal of irrelevant information from working memory: sometimes fast, sometimes slow, and sometimes not at all: Removal of irrelevant information from working memory. *Annals of the New York Academy of Sciences*, 1424(1):239–255, 2018. ISSN 00778923. doi: 10.1111/nyas.13603.

[53] W. E. Hick. On the Rate of Gain of Information. *Quarterly Journal of Experimental Psychology*, 4(1):11–26, 1952. ISSN 0033-555X. doi: 10.1080/17470215208416600.

[54] Shui-I Shih and George Sperling. Measuring and modeling the trajectory of visual spatial attention. *Psychological Review*, 109(2):260–305, 2002. ISSN 1939-1471, 0033-295X. doi: 10.1037/0033-295X.109.2.260.

[55] Flora Bouchacourt and Timothy J. Buschman. A Flexible Model of Working Memory. *Neuron*, 103(1):147–160.e8, 2019. ISSN 08966273. doi: 10.1016/j.neuron.2019.04.020.

[56] 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, 2022. ISSN 00100285. doi: 10.1016/j.cogpsych.2022.101493.

[57] Ivan Tomić, Zahara Girones, Máté Lengyel, and Paul M. Bays. Effects of retention interval on stimulus-specific biases in visual working memory. preprint, in prep.

[58] Rainer Doost and M. T. Turvey. Iconic memory and central processing capacity. *Perception & Psychophysics*, 9(3):269–274, 1971. ISSN 0031-5117, 1532-5962. doi: 10.3758/BF03212646.

[59] Onur Ozan Koyluoglu, Yoni Pertzov, Sanjay Manohar, Masud Husain, and Ila R Fiete. Fundamental bound on the persistence and capacity of short-term memory stored as graded persistent activity. *eLife*, 6, 2017. ISSN 2050-084X. doi: 10.7554/eLife.22225.

[60] George Sperling. A Brief Overview of Computational Models of Spatial, Temporal, and Feature Visual Attention. In Thomas Lachmann and Tina Weis, editors, *Invariances in Human Information Processing*, pages 143–182. Routledge, New York, 2018.

[61] Martin Boerlin and Sophie Denève. Spike-Based Population Coding and Working Memory. *PLoS Computational Biology*, 7(2):e1001080, 2011. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1001080.

[62] Jeffrey M. Beck, Wei Ji Ma, Roozbeh Kiani, Tim Hanks, Anne K. Churchland, Jamie Roitman, Michael N. Shadlen, Peter E. Latham, and Alexandre Pouget. Probabilistic Population Codes for Bayesian Decision Making. *Neuron*, 60(6):1142–1152, 2008. ISSN 08966273. doi: 10.1016/j.neuron.2008.09.021.

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[63] Geoffrey R. Loftus, Janine Duncan, and Paul Gehrig. On the time course of perceptual information that results from a brief visual presentation. *Journal of Experimental Psychology: Human Perception and Performance*, 18(2):530–549, 1992. ISSN 1939-1277, 0096-1523. doi: 10.1037/0096-1523.18.2.530.

[64] Z.-L. Lu, J. Neuse, S. Madigan, and B. A. Dosher. Fast decay of iconic memory in observers with mild cognitive impairments. *Proceedings of the National Academy of Sciences*, 102(5): 1797–1802, 2005. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0408402102.

[65] Timothy J. Ricker, Lauren R. Spiegel, and Nelson Cowan. Time-based loss in visual short-term memory is from trace decay, not temporal distinctiveness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(6):1510–1523, 2014. ISSN 1939-1285, 0278-7393. doi: 10.1037/xlm0000018.

[66] Hongsup Shin, Qijia Zou, and Wei Ji Ma. The effects of delay duration on visual working memory for orientation. *Journal of Vision*, 17(14):10, 2017. ISSN 1534-7362. doi: 10.1167/17.14.10.

[67] Weiwei Zhang and Steven J. Luck. Sudden Death and Gradual Decay in Visual Working Memory. *Psychological Science*, 20(4):423–428, 2009. ISSN 0956-7976, 1467-9280. doi: 10.1111/j.1467-9280.2009.02322.x.

[68] Carlos D Brody, Ranulfo Romo, and Adam Kepecs. Basic mechanisms for graded persistent activity: discrete attractors, continuous attractors, and dynamic representations. *Current Opinion in Neurobiology*, 13(2):204–211, 2003. ISSN 09594388. doi: 10.1016/S0959-4388(03)00050-3.

[69] Mikail Khona and Ila R. Fiete. Attractor and integrator networks in the brain. *Nature Reviews Neuroscience*, 23(12):744–766, 2022. ISSN 1471-003X, 1471-0048. doi: 10.1038/s41583-022-00642-0.

[70] Sanjeev B. Khanna, Adam C. Snyder, and Matthew A. Smith. Distinct Sources of Variability Affect Eye Movement Preparation. *The Journal of Neuroscience*, 39(23):4511–4526, 2019. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.2329-18.2019.

[71] M. H. Chang, K. M. Armstrong, and T. Moore. Dissociation of Response Variability from Firing Rate Effects in Frontal Eye Field Neurons during Visual Stimulation, Working Memory, and Attention. *Journal of Neuroscience*, 32(6):2204–2216, 2012. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.2967-11.2012.

[72] Martin Rolfs and Richard Schweitzer. Coupling perception to action through incidental sensory consequences of motor behaviour. *Nature Reviews Psychology*, 1(2):112–123, 2022. ISSN 2731-0574. doi: 10.1038/s44159-021-00015-x.

[73] Claus Bundesen. A theory of visual attention. *Psychological Review*, 97(4):523–547, 1990. ISSN 1939-1471, 0033-295X. doi: 10.1037/0033-295X.97.4.523.

[74] George Sperling and Erich Weichselgartner. Episodic theory of the dynamics of spatial attention. *Psychological Review*, 102(3):503–532, 1995. ISSN 1939-1471, 0033-295X. doi: 10.1037/0033-295X.102.3.503.

[75] Philip L. Smith and Roger Ratcliff. An integrated theory of attention and decision making in visual signal detection. *Psychological Review*, 116(2):283–317, 2009. ISSN 1939-1471, 0033-295X. doi: 10.1037/a0015156.

[76] Klaus Oberauer and Reinhold Kliegl. A formal model of capacity limits in working memory. *Journal of Memory and Language*, 55(4):601–626, 2006. ISSN 0749596X. doi: 10.1016/j.jml.2006.08.009.

[77] M. Usher and J. L. McClelland. The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*, 108(3):550–592, 2001. ISSN 0033-295X. doi: 10.1037/0033-295x.108.3.550.

[78] David K. Sewell, Simon D. Lilburn, and Philip L. Smith. An information capacity limitation of visual short-term memory. *Journal of Experimental Psychology: Human Perception and Performance*, 40(6):2214–2242, 2014. ISSN 1939-1277, 0096-1523. doi: 10.1037/a0037744.

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[79] T. J. Buschman, M. Siegel, J. E. Roy, and E. K. Miller. Neural substrates of cognitive capacity limitations. *Proceedings of the National Academy of Sciences*, 108(27):11252–11255, 2011. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1104666108.

[80] Thomas C. Sprague, Edward F. Ester, and John T. Serences. Reconstructions of Information in Visual Spatial Working Memory Degrade with Memory Load. *Current Biology*, 24(18): 2174–2180, 2014. ISSN 09609822. doi: 10.1016/j.cub.2014.07.066.

[81] Robert Taylor and Paul M. Bays. Theory of neural coding predicts an upper bound on estimates of memory variability. *Psychological Review*, 127(5):700–718, 2020. ISSN 1939-1471, 0033-295X. doi: 10.1037/rev0000189.

[82] Nikolaus Kriegeskorte and Xue-Xin Wei. Neural tuning and representational geometry. *Nature Reviews Neuroscience*, pages 1–16, 2021. ISSN 1471-0048. doi: 10.1038/s41583-021-00502-3.

[83] Ariel Zylberberg, Stanislas Dehaene, Gabriel B. Mindlin, and Mariano Sigman. Neurophysiological bases of exponential sensory decay and top-down memory retrieval: a model. *Frontiers in Computational Neuroscience*, 3, 2009. ISSN 16625188. doi: 10.3389/neuro.10.004.2009.

[84] Qing Yu, Matthew F. Panichello, Ying Cai, Bradley R. Postle, and Timothy J. Buschman. Delay-period activity in frontal, parietal, and occipital cortex tracks noise and biases in visual working memory. *PLOS Biology*, 18(9):e3000854, 2020. ISSN 1545-7885. doi: 10.1371/journal.pbio.3000854.

[85] Rosanne L. Rademaker, Chaipat Chunharas, and John T. Serences. Coexisting representations of sensory and mnemonic information in human visual cortex. *Nature Neuroscience*, 22(8): 1336–1344, 2019. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-019-0428-x.

[86] John T. Serences. Neural mechanisms of information storage in visual short-term memory. *Vision Research*, 128:53–67, 2016. ISSN 00426989. doi: 10.1016/j.visres.2016.09.010.

[87] Yaoda Xu. Reevaluating the Sensory Account of Visual Working Memory Storage. *Trends in Cognitive Sciences*, 2017. ISSN 13646613. doi: 10.1016/j.tics.2017.06.013.

[88] P.S Goldman-Rakic. Cellular basis of working memory. *Neuron*, 14(3):477–485, 1995. ISSN 08966273. doi: 10.1016/0896-6273(95)90304-6.

[89] Katherine C Bettencourt and Yaoda Xu. Decoding the content of visual short-term memory under distraction in occipital and parietal areas. *Nature Neuroscience*, 19(1):150–157, 2016. ISSN 1097-6256, 1546-1726. doi: 10.1038/nn.4174.

[90] Stephenie A. Harrison and Frank Tong. Decoding reveals the contents of visual working memory in early visual areas. *Nature*, 458(7238):632–635, 2009. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature07832.

[91] Thomas B. Christophel, Polina Iamshchinina, Chang Yan, Carsten Allefeld, and John-Dylan Haynes. Cortical specialization for attended versus unattended working memory. *Nature Neuroscience*, 2018. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-018-0094-4.

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

[93] Laura Rerko, Klaus Oberauer, and Hsuan-Yu Lin. Spatial transposition gradients in visual working memory. *The Quarterly Journal of Experimental Psychology*, 67(1):3–15, 2014. ISSN 1747-0218, 1747-0226. doi: 10.1080/17470218.2013.789543.

[94] Eckart Zimmermann, M. Concetta Morrone, and David C. Burr. Spatial position information accumulates steadily over time. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 33(47):18396–18401, 2013. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.1864-13.2013.

[95] Melonie Williams, Sang W. Hong, Min-Suk Kang, Nancy B. Carlisle, and Geoffrey F. Woodman. The benefit of forgetting. *Psychonomic Bulletin & Review*, 20(2):348–355, 2013. ISSN 1069-9384, 1531-5320. doi: 10.3758/s13423-012-0354-3.

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[96] Joshua J. LaRocque, Jarrod A. Lewis-Peacock, Andrew T. Drysdale, Klaus Oberauer, and Bradley R. Postle. Decoding Attended Information in Short-term Memory: An EEG Study. *Journal of Cognitive Neuroscience*, 25(1):127–142, 2013. ISSN 0898-929X, 1530-8898.

[97] Robert Taylor, Ivan Tomić, David Aagten-Murphy, and Paul M. Bays. Working memory is updated by reallocation of resources from obsolete to new items. *Attention, Perception, & Psychophysics*, 2022. ISSN 1943-3921, 1943-393X. doi: 10.3758/s13414-022-02584-2.

[98] D. H. Kelly and E. Martinez-Uriegas. Measurements of chromatic and achromatic afterimages. *Journal of the Optical Society of America. A, Optics and Image Science*, 10(1):29–37, 1993. ISSN 0740-3232.

[99] Naotsugu Tsuchiya and Christof Koch. Continuous flash suppression reduces negative afterimages. *Nature Neuroscience*, 8(8):1096–1101, 2005. ISSN 1097-6256, 1546-1726. doi: 10.1038/nn1500.

[100] Cajo J. F. Ter Braak. A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces. *Statistics and Computing*, 16 (3):239–249, 2006. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-006-8769-1.

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary information

Methods

Participants

A total of twenty-three naive observers (12 females, 11 males; aged 18–34) took part in the study after giving informed consent in accordance with the Declaration of Helsinki. Ten observers participated in Experiment 1 and thirteen observers participated in Experiment 2. Volunteers were recruited through the Cambridge Psychology research sign-up system. All observers reported normal color vision and normal or corrected-to-normal visual acuity, and were remunerated £10/hr for their participation.

General methods

Experimental setup. Stimuli were presented on a 69 cm gamma-corrected LCD monitor with a refresh rate of 144 Hz. Participants were seated in a dark room and viewed the monitor at a distance of 60 cm, with their head supported by a forehead and chin rest. Responses were collected using Magic Trackpad 2, a pointing device (16 x 11.5 cm) with a tactile sensor operating at ~90 Hz (Apple Inc.). Eye position was monitored online at 1000 Hz using an infrared eye tracker (SR Research). Stimulus presentation and response registration were controlled by a script written in Psychtoolbox and run using Matlab (The Mathworks Inc.).

Stimuli. Memory stimuli consisted of randomly oriented Gabor patches (wavelength of the sinusoid, 0.65° of visual angle; s.d. of Gaussian envelope, 0.5°) presented on a uniform mid-grey background. The contrast of Gabor patches varied between experiments (see below). Memory stimulus positions were randomly chosen from a set of ten equidistant locations on the perimeter of an invisible circle with radius 6° centered at fixation. At the start of each trial, a black fixation annulus was shown ($r = 0.15°$ and $R = 0.25°$) in the display center. Once steady fixation was registered, the size of the inner radius increased ($r = 0.2°$). Observers perceived this change as the annulus becoming thinner. The fixation annulus then stayed visible throughout the trial. Items were cued for recall by displaying a black arrow (2° length) extending from the center of the display and pointing to one of the previously occupied locations without overlapping with it.

Procedure. Each trial started with presentation of the central fixation annulus. Observers were required to maintain gaze fixation for 500 ms within a radius of 2° around the central annulus in order for a trial to proceed. Following stable fixation, the appearance of the fixation annulus changed, indicating that the memory array would appear in 500 ms. The memory sample array consisting of 1, 4, or 10 randomly oriented Gabor patches was then presented. This was followed by a delay period and finally a cue display, indicating to observers to report the memorized orientation of an item previously displayed at the indicated location.

Observers were instructed to reproduce the remembered orientation as accurately and as quickly as possible by executing a single movement of their index fingertip over the surface of the touchpad located centrally in front of them. Simultaneously with the observer's movement, a blue line appeared on the screen, extending from the center of the screen and mimicking the observer's response in real-time. The response was terminated if one of the following conditions was satisfied: the observer stopped movement for 500 ms; the observer lifted their finger from the touchpad; or the response line reached the edge of the display. This was followed by a feedback display, consisting of the actual orientation (shown with a white line) and reported orientation (shown with a blue line) overlaid at the location of the cued item. The recalled orientation was calculated as the angle of the line connecting a starting point and an endpoint of hand movement on the touchpad.

Observers were required to maintain central fixation during the stimulus presentation and delay phase. If gaze position deviated by more than 2° a message appeared on the screen, and the trial was aborted and restarted with newly randomized orientations. Participants completed the task in blocks of 50 trials, and each block corresponded to one experimental condition. The order of blocks was randomized for every observer. At the beginning of the testing session observers familiarized themselves with the task and experimental setup by doing at most 50 practice trials.

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Experiment 1

In Experiment 1 we investigated the temporal dynamics of VWM fidelity over short delays by presenting observers with sets of stimuli of variable size and then cueing one of them for recall after a variable delay relative to the stimuli offset. A typical trial sequence is shown in Figure S1A. The memory sample array (Michelson contrast = 0.5) was presented for 200 ms. In 50% of trials, the stimuli changed phase (by 180°) and contrast (Michelson contrast = 1) for the last 50 ms of presentation, while remaining at the same orientation. This manipulation was intended to minimize retinal after-effects (see e.g. 98 for similar techniques). The stimuli offset was followed by a variable blank delay of 0, 100, 200, 400, or 1000 ms, after which one item was cued for recall. In one additional condition, the cue was instead presented simultaneously with the memory sample array, indicating an item while it was still visible on the screen (Fig. S1B).

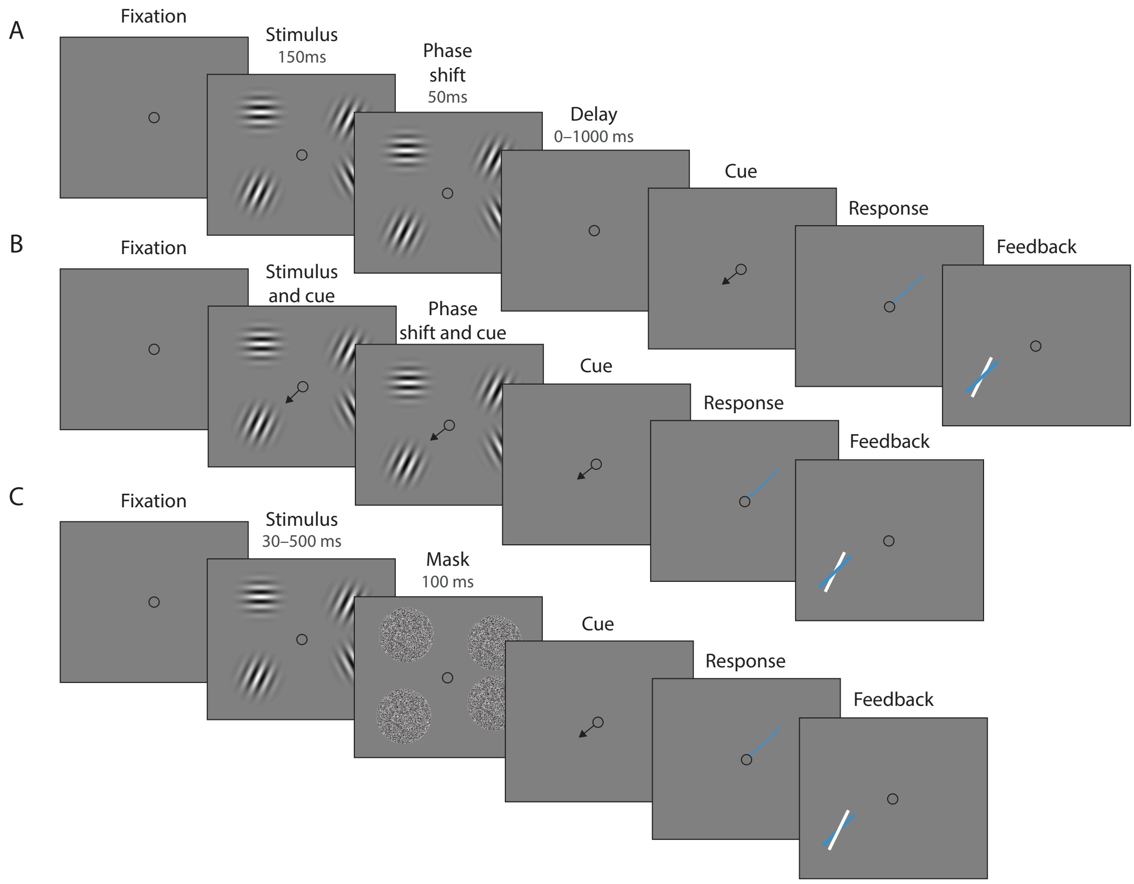

Figure S1: Experimental procedure. (**A**) Experiment 1. On each trial, a memory array was presented consisting of 1, 4, or 10 randomly oriented Gabor stimuli. In 50% of all trials, the stimuli underwent a change of phase and contrast towards the end of the exposure period intended to minimize retinal aftereffects. After a variable delay, an arrow cue was shown pointing towards the location of one stimulus from the preceding array. Observers reported the remembered orientation of the cued stimulus by swiping their index finger on the touchpad. The response was followed by feedback showing the true orientation. (**B**) In a proportion of trials, the cue was presented simultaneously with the stimuli. (**C**) Experiment 2. On each trial a memory array consisting of 1, 4, or 10 randomly oriented Gabors was presented for a variable duration, and followed by a white noise flickering mask. The mask was replaced by an arrow cue pointing towards the location of one stimulus from the preceding array. Observers reported its remembered its orientation and received feedback as in Experiment 1. Stimuli are not drawn to scale.

Each observer completed a total of 1800 trials, split into 36 blocks. The experiment was organized such that half of the observers first completed 18 blocks with phase shift (see above), and the other half first completed blocks without phase shift. Except for this constraint, block order was

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

randomized for every observer. The testing was divided into four equal testing sessions, each lasting approximately 1.5 hours, with a separation of at least one day between sessions.

Experiment 2

In Experiment 2 we investigated the temporal dynamics of VWM fidelity during encoding. To this end, we displayed oriented stimuli for a variable duration and in sets of variable size. The experiment was similar to the previous experiment with a few exceptions (Fig. S1C). Each trial started with a presentation of a fixation annulus, followed by a memory array (Michelson contrast = 0.3). The stimuli stayed on the screen for a variable duration of 30, 48, 77, 122, 196, 313, or 500 ms, and were then replaced by noise masks (100 ms). Mask stimuli consisted of white noise at full contrast, windowed with a Gaussian envelope (0.5° s.d.) and flickering at 35 Hz. At the offset of the masking stimuli, one memory item was cued for recall. Each observer completed 21 blocks, for a total of 1050 trials. Blocks were spread over two testing sessions, each lasting approximately 1.5 hours, and taking place on different days. Observers completed 10 blocks in the first, and the remaining 11 blocks in the second session.

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Minimizing retinal after-effects

We assessed the method of minimizing retinal afterimages by repeating all measurements, with the exception of not using phase shift of stimuli (Fig S1A). We predicted retinal afterimages could serve as an additional source of information, but only for a brief period after stimuli offset. Therefore, here we expected to see better performance for brief delays compared to conditions with phase shift. Figure S2A shows recall error increased with both set size and delay. Both of these effects were statistically significant, as well as their interaction (set size: $F_{(2,18)} = 47.3, p < .001, \eta^2 = .31$; delay time: $F_{(5,45)} = 48.4, p < .001, \eta^2 = .26$; interaction: $F_{(10,90)} = 21.3, p < .001, \eta^2 = .14$), reminiscent of findings for data with phase shift.

Next, we focused on the comparison of conditions with and without phase shift of stimuli (Fig S2B). We illustrate the difference in performance by subtracting RMSE obtained in the condition without phase shift (Fig 3B) from RMSE shown in Figure S2A. Negative values indicate better performance in a condition without phase shift. As predicted, the overall pattern of data suggested performance was comparable for 1 item across all delays, and for all set sizes for extreme delays (simultaneous presentation and 1000 ms), indicated by the difference values around 0. We confirmed the difference in recall error for 1 item across all delays did not differ consistently with and without phase shift, as neither phase shift ($F_{(1,9)} = 0.03, p = .86, \eta^2 < .001, BF_{incl} = 0.143$) nor the interaction of phase shift and delay ($F_{(5,45)} = 0.41, p = .89, \eta^2 = .00, BF_{incl} = 0.042$) reached significance. Based on this result, we conducted all remaining analyses using only the remaining two set sizes. We ran separate repeated measures ANOVAs for each delay using phase shift and set size as factors. The pattern of results we observed was clear: performance was comparable with and without phase shift with the simultaneous presentation and 1000 ms delay (phase shift, $F_{(1,9)} \leq 1.08, p \geq .33, \eta^2 \leq .002, BF_{excl} \geq 3.62$; interaction, $F_{(2,18)} \leq 0.8, p \geq .44, \eta^2 \leq .02, BF_{excl} \geq 3.39$), while for the remaining intermediate delays recall error was consistently lower when phase shift was omitted (phase shift, $F_{(1,9)} \geq 5.8, p \leq .039, \eta^2 \geq .06$; interaction, $F_{(1,9)} \leq 2.8, p \geq .13, \eta^2 \leq .001$).

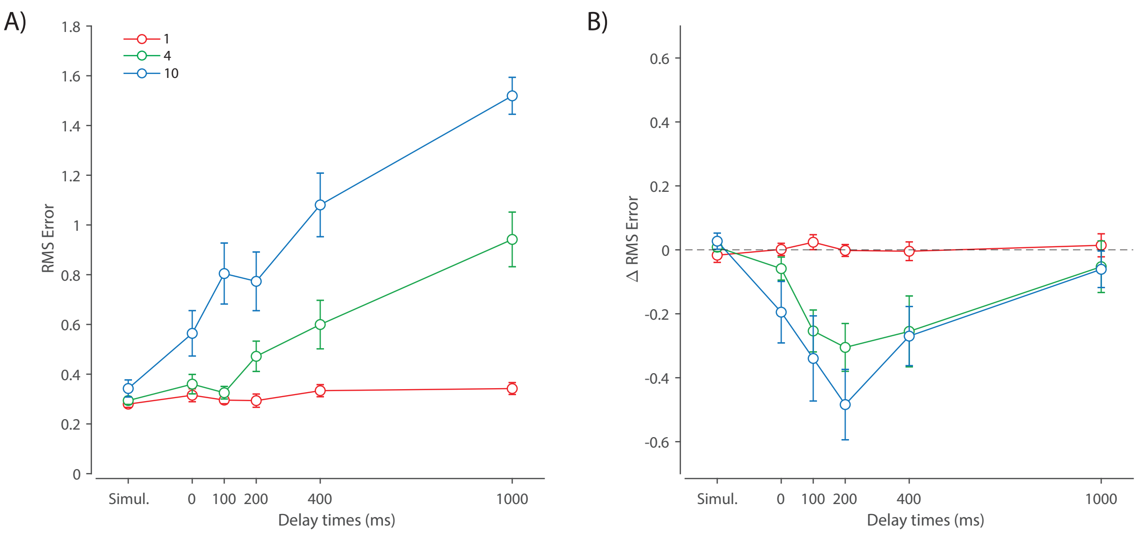

Figure S2: (A) Experiment 1 RMSE for trials without phase shift. (B) Differences in RMSE between trials with and without phase shift across set size and delay conditions. Negative values indicate better performance in the condition without phase shift.

Taken together, performance with and without phase shift of stimuli was comparable in perceptual condition (simultaneous presentation) and with the longest delay, suggesting phase shift did not change visibility or encoding of information into VWM. In contrast, we found strong evidence that observers had access to an additional source of information over intermediate delays when phase shift was not used, demonstrated by a better recall performance from 0 ms to 400 ms delay. Specifically, this source of information was available immediately after stimuli offset and was short-lived, consistent with the theoretical description of retinal afterimages [99].

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1129 ## Swap error estimates

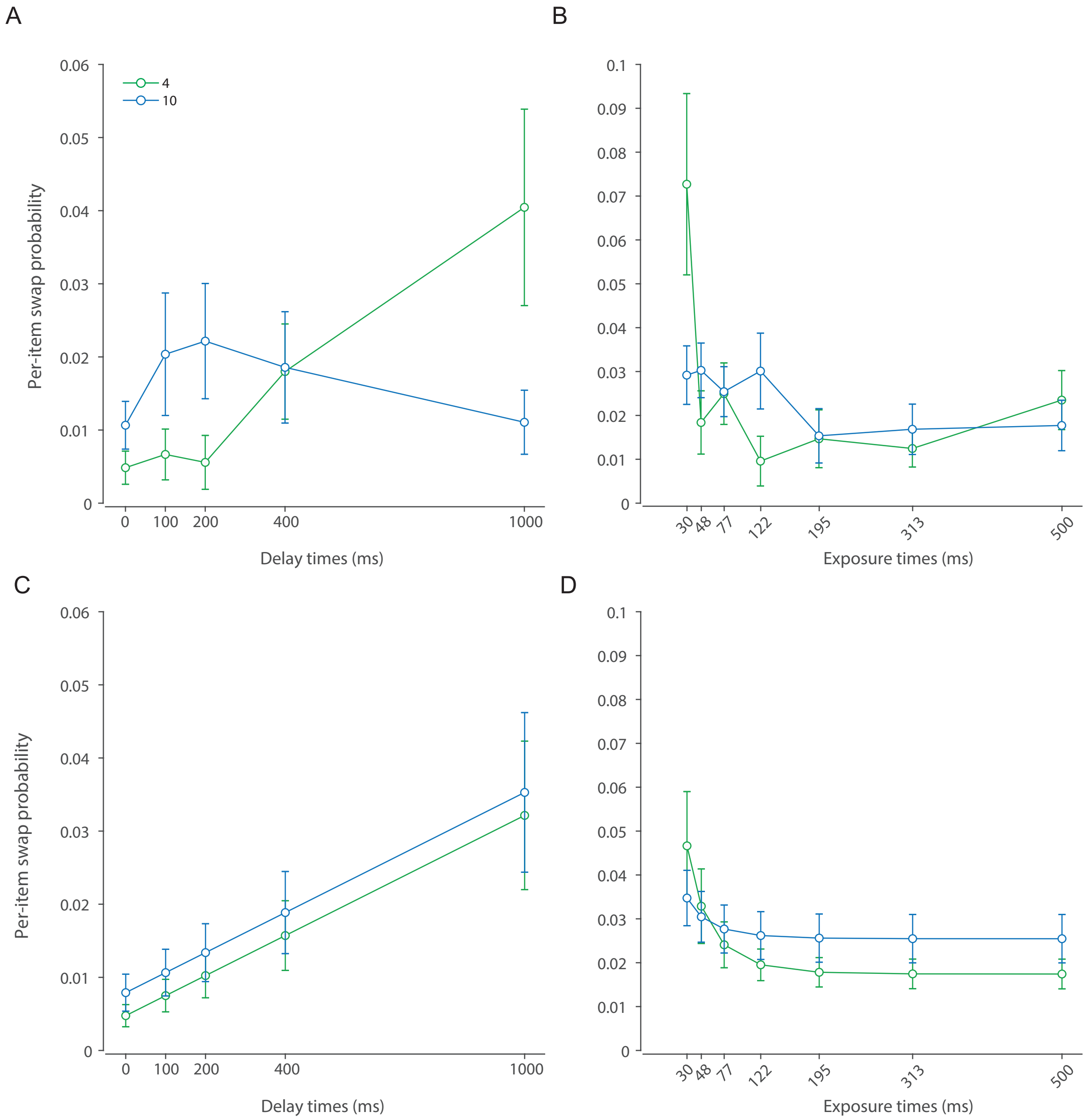

Figure S3: Swap error estimates. (**A&B**) Probability of swap errors estimated from empirical data using the three-component mixture model [92] in Experiment 1 (A) and Experiment 2 (B). (**C&D**) Probability of swap errors in best-fitting DyNR model in Experiment 1 (C) and Experiment 2 (D).

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1130 # Alternative models' fits

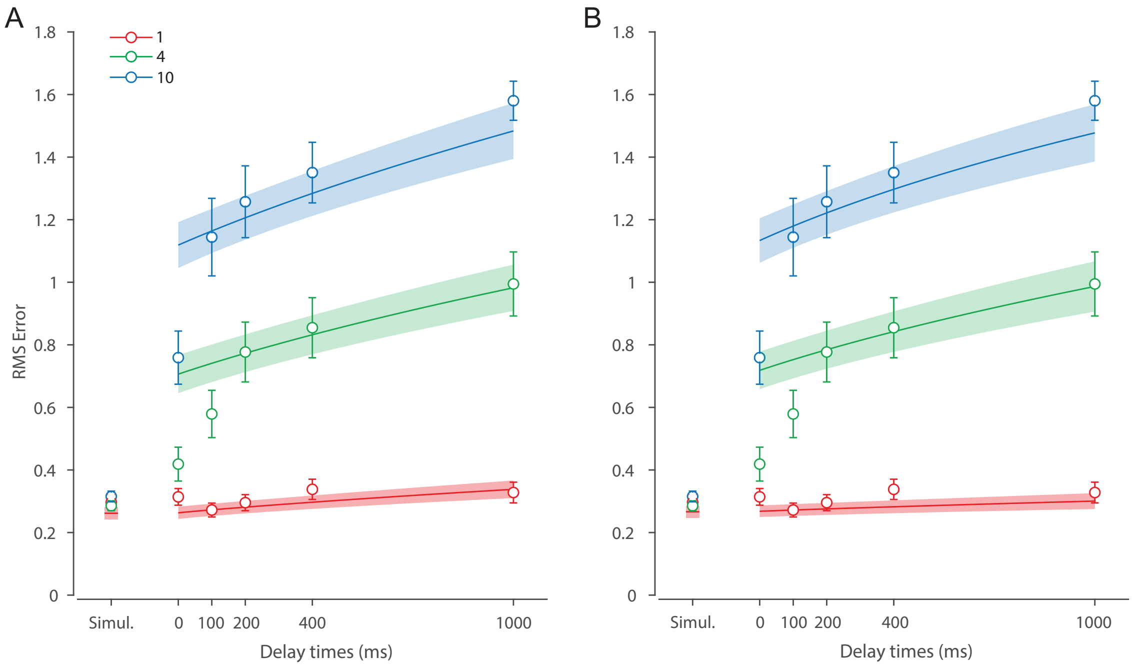

Figure S4: Experiment 1 behavioral data and model fit for the DyNR model without sensory persistence after stimulus offset. (**A**) A version of the DyNR model with equal diffusion across set sizes. (**B**) A version of the DyNR model with diffusion that scales with set size.

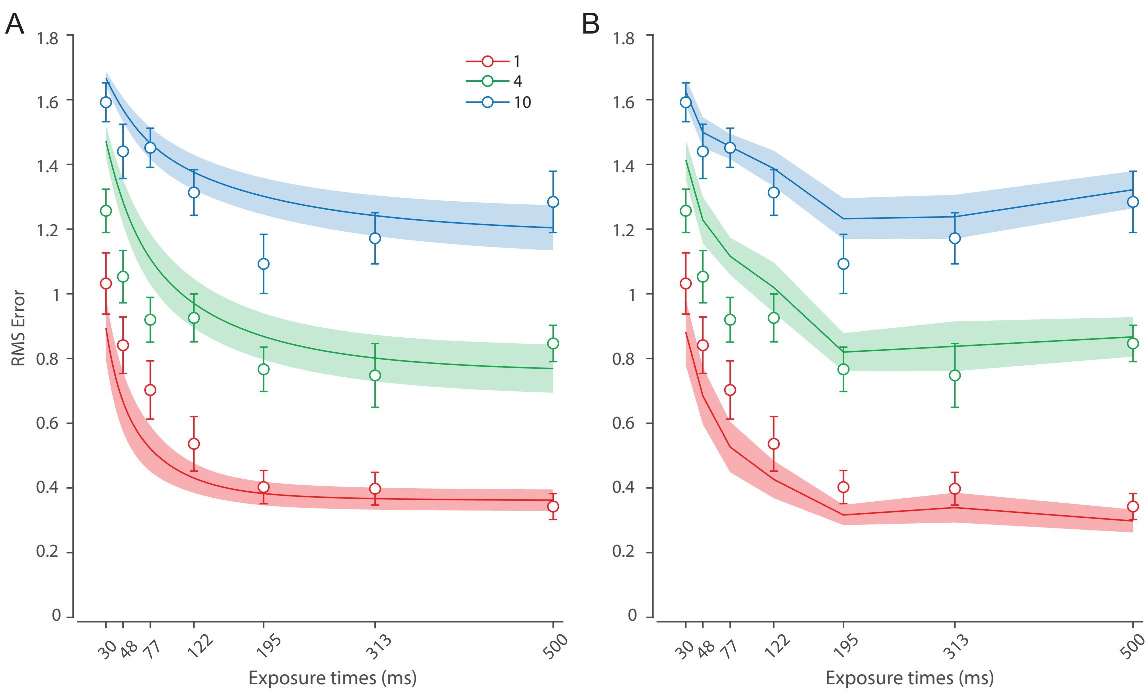

Figure S5: Experiment 2 behavioral data and model fit for the neural model without sensory persistence after stimulus offset. (**A**) A version of the DyNR model without sensory persistence. (**B**) Separate fits of the simplified neural model to each exposure time.

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

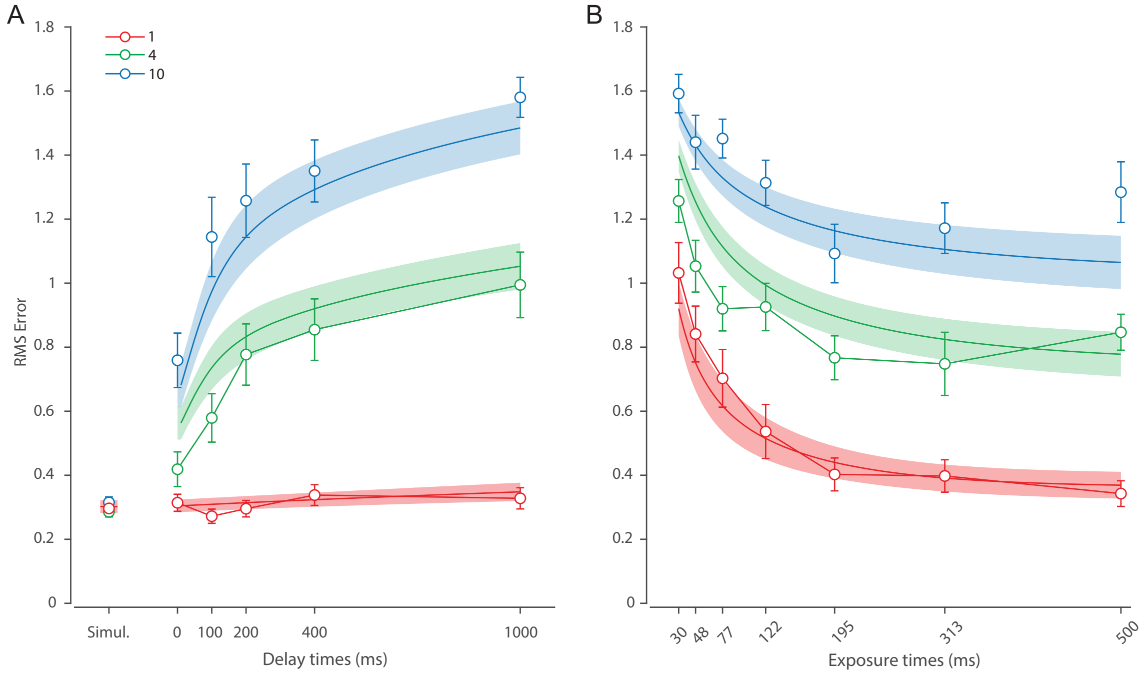

Figure S6: Behavioral data and model fit for the DyNR model without the cue processing time for (**A**) Experiment 1 and (**B**) Experiment 2.

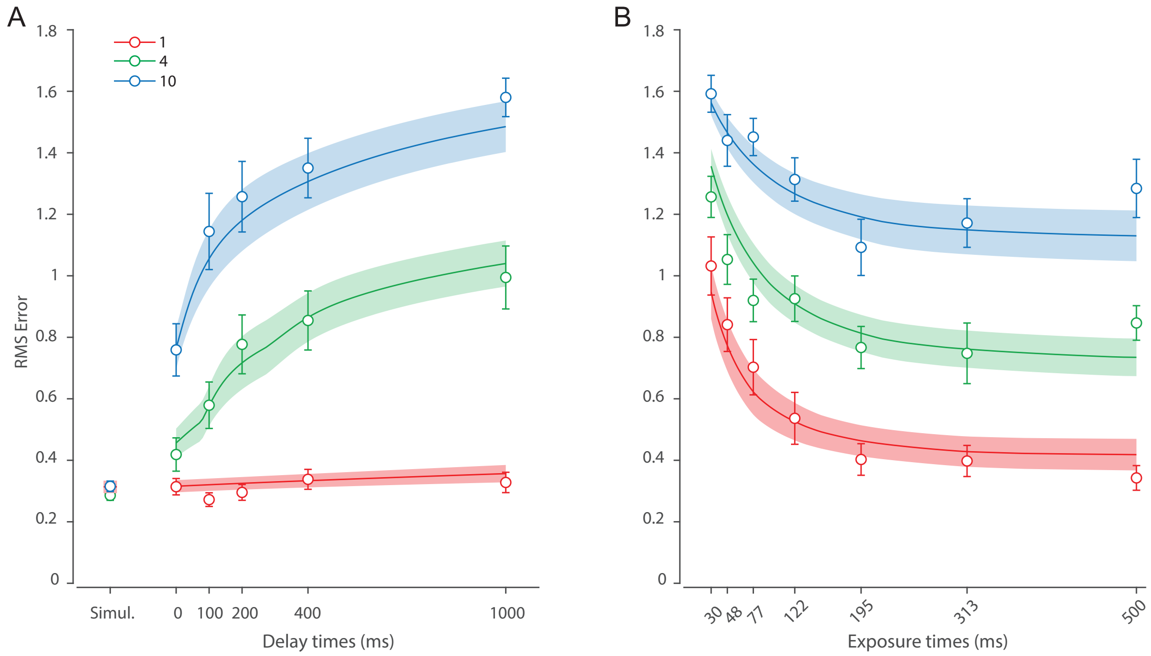

Figure S7: Behavioral data and model fit for a neural model with constant accumulation of information into WM for (**A**) Experiment 1 and (**B**) Experiment 2.

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

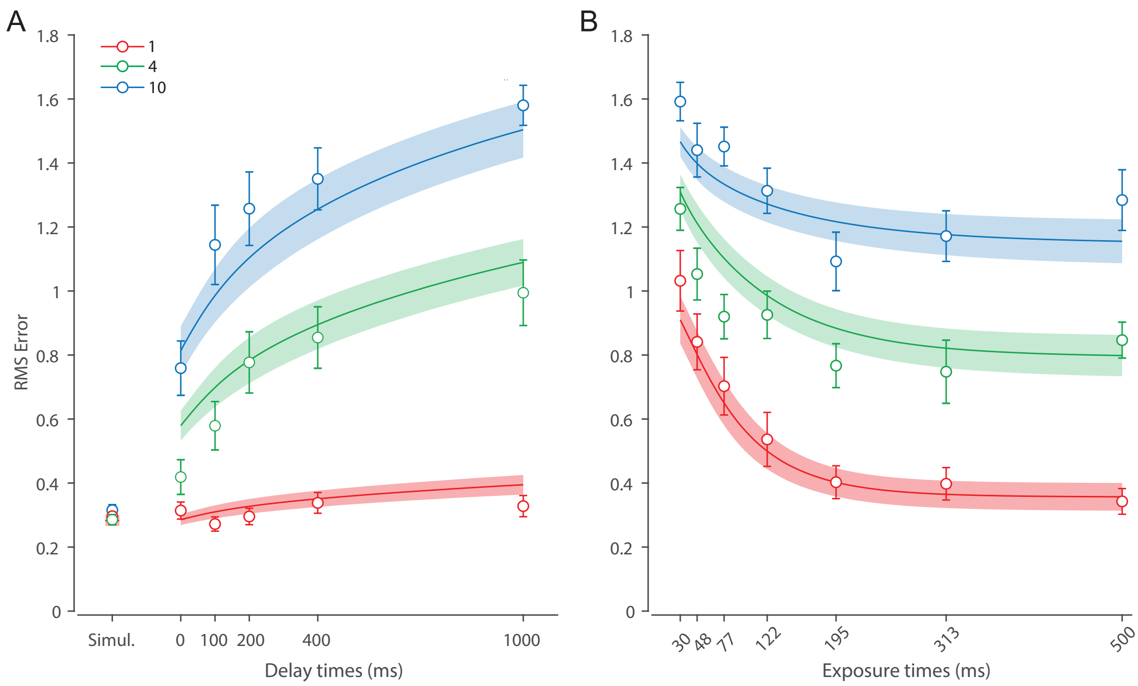

Figure S8: Behavioral data and model fit for a neural model with the direct read-out of information from sensory memory for (**A**) Experiment 1 and (**B**) Experiment 2.

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Additional dataset 1

To further investigate the role of diffusion in memory dynamics, we analysed an additional dataset collected in our lab [57]. In this experiment we varied the set size and delay duration similar to Experiment 1. In contrast to Experiment 1, we used longer memory delays, which allowed us to examine the diffusion mechanism on a more suitable time scale. Moreover, memory delays used in this study are out of reach of the decaying sensory information, enabling us to investigate the diffusion without changes in the neural signal strength post-cue.

Methods

Ten observers (6 females, 4 males, aged 18-34) took part in this experiment. The data for this experiment was collected using the same equipment and the testing setting as described for the main experiments. A typical trial sequence is illustrated in Fig. S9. Each trial began with the presentation of a central annulus which served as a fixation point. Once a stable fixation was achieved, the inner annulus radius changed indicating that stimuli would appear in 500 ms. The memory sample array was then presented for a duration of 500 ms. The array consisted of one or three randomly oriented black bars (length 2.8°). Each bar was positioned in one of six predetermined locations equally distributed around the circle with a radius of 5° around center of the screen. Each bar was presented along with a placeholder circle (radius 1.5°).

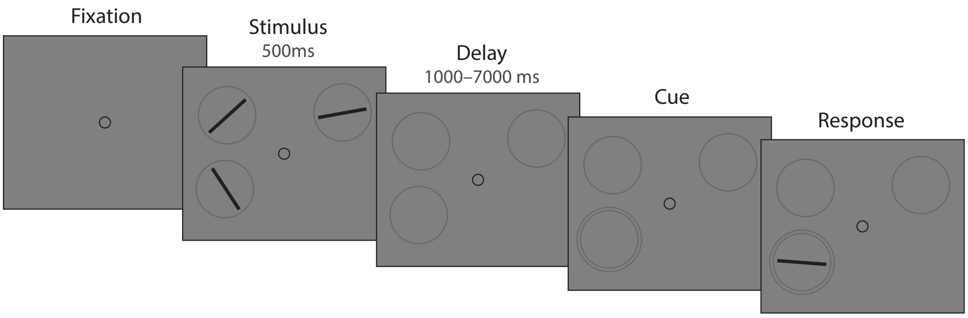

Figure S9: Experimental procedure. Stimuli are not drawn to scale.

Memory array presentation was followed by a memory delay during which fixation circle and placeholders stayed visible. The retention interval was either 1 or 7 seconds long. After that, one stimulus was randomly cued for recall. The cue consisted of a second, larger circle drawn around one of the placeholders. Observers were instructed to start rotating a response dial (Griffin Technology PowerMate USB) once they were ready to respond. After the rotation of the response dial was detected, a randomly oriented black bar was displayed within the placeholder. Observers were instructed to rotate the dial until the displayed bar matched the remembered orientation of the cued item. Observers confirmed their response by pressing the dial. Trials with different set sizes and delay durations were randomly interleaved.

Eye movements were monitored from the beginning of the trial until stimuli offset, and observers were required to hold steady fixation during that period. If the gaze position deviated by more than 2° a message appeared on the screen and the trial was aborted and restarted with new orientations. Each observer completed 700 trials, divided into two sessions and each consisting of 7 equal blocks. Two sessions were separated by at least one day, and each lasted approximately 1 hour. At the beginning of each session observers familiarized themselves with the task and experimental setup by doing at most 50 practice trials.

Results

Behavioral data. Recall performance is shown in Figure S10. As predicted, response error increased with set size and memory delay. A repeated measures ANOVA revealed a significant effect of set size ($F_{(1,9)} = 111.17, p < .001, \eta^2 = .76$) and memory interval ($F_{(1,9)} = 58.14, p < .001, \eta^2 = .12$),

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and their interaction ($F_{(1,9)} = 10.66, p = .01, \eta^2 = .02$) on response error. Moreover, conducting paired t-tests within each set size revealed recall error increased with the delay with set size 1 ($t_{(9)} = 5.83, p < .001, d = 1.84$) and set size 3 ($t_{(9)} = 5.78, p < .001, d = 1.83$). The interaction effect was a consequence of a larger increase in error with delay for set size 3 compared to set size 1 ($\Delta\text{RMSE} = \text{RMSE}_{7000ms} - \text{RMSE}_{1000ms}$; $t_{(9)} = 3.27, p = .01, d = 1.03$). These results are consistent with Experiment 1, corroborating our finding that increasing the set size and delay time have a disadvantageous effect on memory fidelity.

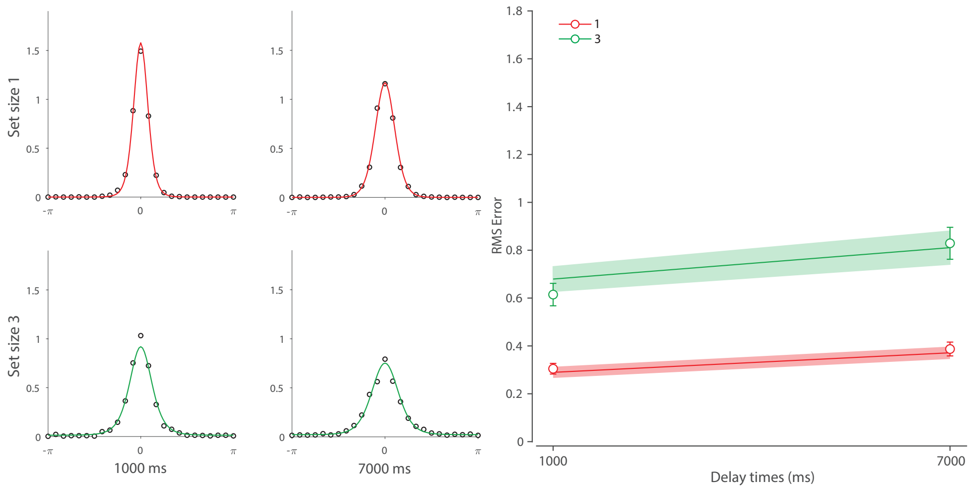

Figure S10: Behavioral data and model fit for Experiment 1a

Neural model. We fitted the DyNR model to the data to test whether noise-driven diffusion is sufficient to account for changes in recall fidelity with longer memory intervals. We applied a simplified version of the model without sensory decay and VWM accumulation components. This was justified given that estimate of sensory decay from Experiment 1 was shorter (mean life $\tau = 0.21$) than the shortest interval used in this experiment (1 s). Moreover, based on our findings in Experiment 2, we argue that a display duration of 500 ms is sufficient to fully encode objects into VWM.

Curves in Figure S10 show fits of the model with maximum likelihood (ML) parameters (mean ± se: population gain $\gamma = 385.02 \pm 208.3$, tuning width $\kappa = 2.67 \pm 0.43$, cue processing constant $b = 0.68 \pm .67$, base diffusion $\sigma^2_{\text{diff}} = 0.009 \pm 0.001$, swap probability $p = 0.005 \pm 0.002$). The model provided an excellent quantitative fit to response distributions and summary statistics (Fig. S10), successfully explaining the adverse effects of set size and memory interval on recall fidelity. Critically, and consistent with results from Experiment 1, the proposed DyNR model provided a better fit to human response error compared to the matching model without diffusion ($\Delta\text{AIC} = 144.75$) or the model in which diffusion rate increases with set size ($\Delta\text{AIC} = 42.3$). In conclusion, this result shows that variability in representations over longer memory intervals can be fully accounted for by noise-driven accumulation without changes in the representational signal [9, 10, 26].

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Additional dataset 2

To further validate predictions of the DyNR model we fitted it to an existing working memory study (Experiment 1 in 35). This study focused on the role of temporal dynamics during WM encoding, thereby addressing the same question as our Experiment 2. In contrast to our Experiment 2, Bays et al. [35] used a longer delay period (1100 ms), precluding the strengthening influence of decaying sensory information on recall. This dataset therefore isolates the initial information accumulation process during stimuli presentation.

Methods

The observers (N = 32) performed a continuous report task in which a variable number of oriented bars was presented for a variable duration, followed by a pattern mask (100 ms) and a 1-second delay period after which one of the items was probed for recall. Set size was manipulated between observers and exposure duration was manipulated within observers. Each observer performed 100 trials per exposure duration, for a total of 25600 trials in the study. A more detailed description of the experiment is provided in Bays et al. [35].

Analysis

Considering only exposure duration in this experiment was manipulated at the observer level, we decided to expand our modelling approach by employing a Bayesian hierarchical method as a compromise between fitting the data for each observer (i.e., set size) independently and pooling the data across all observers. Using a Bayesian hierarchical modelling, individual-observer parameters are considered samples from population distributions, whose means and variances are estimated based on all available data. In general, this approach has a desirable characteristic of constraining individual-level parameters with the population-level distribution and producing meaningful parameter estimates when a model is fitted across separate groups. The dynamic neural model fitted to the data is identical to the model fitted in Experiment 2, with the exception that here we assumed any existing post-stimulus sensory activity completely diminished by the time of the cue (1100 ms post-stimulus offset), and therefore we did not model sensory decay here. To obtain the hierarchical fit, we used the Differential Evolution Markov Chain algorithm [100]. All individual-level parameters were samples drawn from normal (i.e., Gaussian) distributions, with corresponding mean and standard deviation being constrained by uniform hyperprior distributions. We collected 240000 post-warmup samples across 12 chains and computed median and 95% equal-tailed intervals (ETI) of posterior distributions to obtain the group and individual-level parameter estimates. Prior specifications and empirical data for all analyses can be found along with the published code.

Results

Figure S11 and Figure S12 show empirical distributions and summary statistics across all conditions. Similar to Experiment 2, increasing the exposure duration ($F_{(7,196)} = 110.9, p < .001, \eta^2 = .188$) and decreasing the set size ($F_{(3,28)} = 22.83, p < .001, \eta^2 = .53$) had beneficial effect on response error. Interaction of exposure duration and set size was significant ($F_{(21,196)} = 3.13, p < .001, \eta^2 = .02$). Critically, the pattern of memory fidelity dynamics largely matches the pattern observed in Experiment 2, with response errors decreasing rapidly as presentation duration was increased from the minimum duration, saturating at longer durations. This pattern was consistent across all set sizes, which only differed in the absolute error.

These dynamics were accurately predicted by the DyNR model, both at the level of response distributions (curves in Fig. S11) and summary statistics (curves in Fig. S12). The parameters used to generate model predictions were obtained by taking the individual observer's posterior medians. We observed the following hyperparameters (median and 95% ETI of hyperposterior): population gain $\gamma = 109.47$ (88.1 - 133.57), tuning width $\kappa = 3.23$ (2.6 - 4.03), sensory rise time constant $\tau_{\text{rise}} = 0.0049$ (0.0019 - 0.0091), VWM accumulation time constant $\tau_{\text{WM}} = 0.067 \pm$ (0.051 - 0.087), cue processing constant $b = 0.423$ (0.093 - 0.8436), base diffusion $\sigma^2_{\text{diff}} = 0.095$ (0.057 - 0.149), spatial uncertainty time constant $\tau_{\text{spatial}} = 0.031$ (0.022 - 0.041), swap probability $p = 0.02$ (0.011 - 0.034).

bioRxiv preprint doi: https://doi.org/10.1101/2023.03.27.534406; this version posted March 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

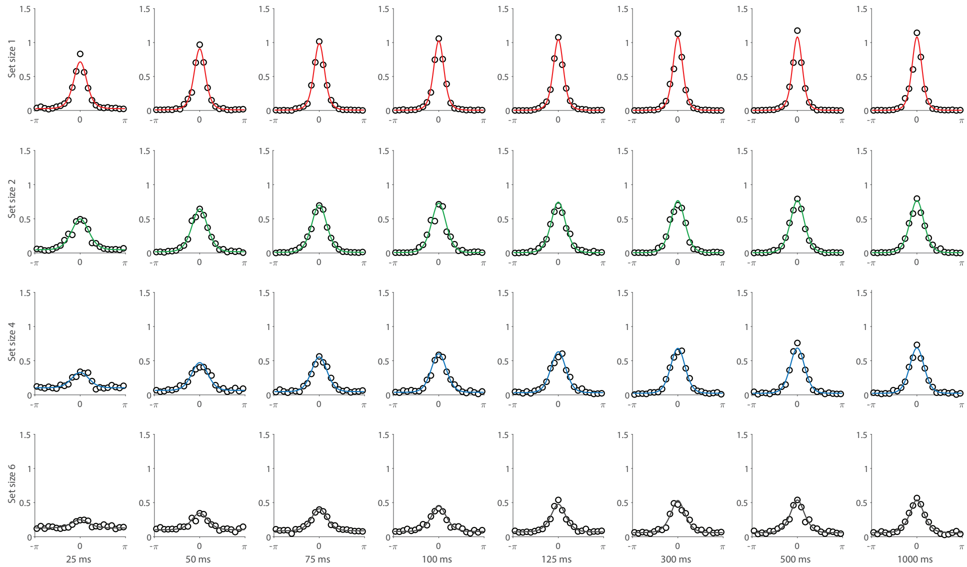

Figure S11: Empirical recall error distributions (black circles) from Experiment 1 in Bays et al. [35] and the DyNR model fits to the data (colored curves).

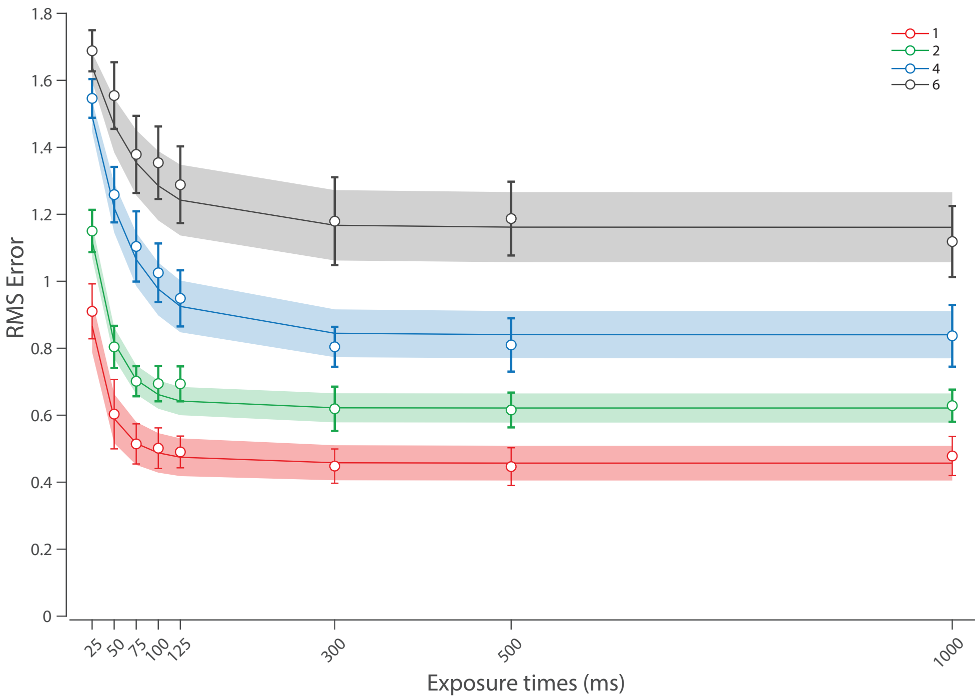

Figure S12: Summary statistics (black circles) from Experiment 1 in Bays et al. [35] and the DyNR model fits to the data (colored curves). The DyNR model was fit to the distributions of recall errors shown in Fig. S11.