www.nature.com/scientificreports

scientific reports

OPEN

Modeling visual working memory using recurrent on-center off-surround neural network with distance dependent inhibition

Rakesh Sengupta[1,2]✉

This paper presents a computational model of visual working memory (VWM) that simulates the processing of spatially distributed objects and their features. The model emphasizes the prioritization of object-related information before feature-related processing, effectively reversing the conventional feedforward order in visual perception. We conduct a detailed stability analysis to demonstrate non-divergence through energy function evaluations, highlighting the robustness of the network under varying input conditions. Additionally, we investigate the model's performance in change detection tasks through error rate analysis, focusing on how receptive field sizes and input configurations (crowded or spaced out) affect accuracy. This paper further extends the distance dependent RNN idea to a hierarchical computational model of visual working memory designed to parallel the structure and function of the visual processing hierarchy (VH). The model introduces a top-down feedback mechanism that dynamically refines object and feature representations, prioritizing object-level clarity initially and progressively localizing feature details through recurrent processing. A novel cross-talk function modulates feature specificity based on feedback depth, capturing the uncertainty in representations before full localization reaches lower visual areas.

Visual working memory (VWM) is a critical cognitive function that allows for the temporary storage and manipulation of visual information. Understanding the neural mechanisms underlying VWM is important for advancing both theoretical neuroscience and practical applications in artificial intelligence (AI). One promising approach to modeling VWM involves the use of recurrent on-center off-surround (OCOS) neural networks, which mimic the inhibitory and excitatory interactions observed in biological neural circuits[1,2]. The current work explores the application of a recurrent OCOS neural network with distance-based inhibition to model VWM, drawing on insights from various studies in the field.

The concept of on-center off-surround inhibition is rooted in the early visual processing mechanisms of the brain, where it plays a crucial role in edge detection and contrast enhancement. Bhatnagar[3] demonstrated the effectiveness of this mechanism in artificial visual cortex models for edge detection, highlighting its potential for more complex visual tasks. In a similar vein, Verma and Sengupta[4] introduced an enumeration module for the human visual system using a recurrent OCOS neural network, showcasing its ability to handle dynamic visual information.

The recurrent nature of these networks allows for the integration of temporal information, which is essential for modeling the continuous and dynamic aspects of VWM. Ross et al.[5] discussed the importance of recurrent interactions in perceptual grouping, emphasizing how these mechanisms contribute to the stability and coherence of visual perception. Similarly, Hansen et al.[6] explored long-range interactions in early vision, providing a framework for understanding how recurrent connections can enhance visual processing.

In addition to recurrent interactions, distance-based inhibition is a key feature of the proposed model. This mechanism ensures that inhibitory signals are modulated based on the spatial distance between neurons, reflecting the spatial organization of neural circuits in the brain. Thielscher and Neumann[7] investigated cortico-cortical interactions in texture boundary detection, illustrating how distance-based inhibition can enhance the detection of complex visual patterns. Maric[8] further explored top-down effects on visual perception, highlighting the role of inhibitory mechanisms in modulating visual processing based on contextual information.

[1]School of Interwoven Arts and Sciences, Krea University, Sector 24, Sri City, Andhra Pradesh 517646, India. [2]Center for Creative Cognition, SR University, Anathasagar, Warangal, Telangana 506371, India. ✉email: qg.rakesh@gmail.com

The integration of these concepts into a single model offers a comprehensive approach to understanding VWM. Kurt[9] characterized surround suppression in motion direction perception, providing insights into how inhibitory mechanisms can enhance motion detection and discrimination. Ogmen[10] examined the spatiotemporal dynamics of visual perception across neural maps and pathways, underscoring the importance of integrating spatial and temporal information in visual processing.

Previously, Sengupta et al.[11] demonstrated that a visual sense of number can emerge from the dynamics of a recurrent OCOS neural network, suggesting that such models can capture higher-order cognitive functions. Thielscher[12] also explored nonlinear recurrent mechanisms for processing surface boundaries, further supporting the potential of recurrent OCOS networks for complex visual tasks. A similar OCOS model has been used to demonstrate certain features (e.w., recency effect) of serial recall in working memory[1], and more recently have been used to look at primacy effects as well[2]. However, such OCOS networks have fully connected single-layer architecture that is a limitation for developing a full account of visual working memory.

In the current paper we aim to advance our understanding of VWM using a recurrent OCOS neural network with distance-based inhibition. By integrating insights from various studies, we propose a model that captures the dynamic and spatial aspects of visual processing, offering a robust framework for future research in both neuroscience and AI.

Advances in our understanding of visual processing have shown the importance of multi-area interactions and top–down feedback in shaping perceptual and memory representations. Neurophysiological and behavioral studies have demonstrated that the visual cortical hierarchy is not merely a feedforward cascade but involves robust re-entrant processing and feedback modulation[13–16]. In particular, work by[15] has highlighted the role of prefrontal cortex in integrating and directing feedback signals, which are essential for resolving ambiguity in sensory inputs and maintaining working memory. These findings provide compelling evidence for a hierarchical organization in which strong feedback pathways play a critical role. Our model capitalizes on these insights by incorporating robust feedback connections in a recurrent on-center off-surround framework, thereby emulating the re-entrant interactions observed in the visual cortical areas. This design choice is aimed at capturing the dynamic interplay between feedforward sensory inputs and top–down control processes, ultimately leading to improved stability and precision in visual working memory representations.

In the following, we begin with a single-layer recurrent OCOS with distance-dependent inhibition and show corresponding network dynamics and its steady state dynamics, that allows us to see how spatial arrangement of stimuli might affect the network output given variation of distance dependent inhibition. Following that, we show how such a network can be expanded into a hierarchical VWM model that allows us to incorporate location and feature based information. We also demonstrate the simulations related to that for varying set sizes.

Distance dependent inhibition on single layer recurrent on-center off-surround network

The network is structured as an additive on-center off-surround recurrent neural network, comprising N nodes organized in an square grid. Each node is connected to a sparse subset of other nodes, facilitating lateral inhibition. The connectivity is determined by a distance metric, such that each node connects only to its neighboring nodes within a specified receptive field size (see Fig. 1).

The lateral inhibition between nodes is modulated based on their distances, defined as:

$$\beta_{ij} = \begin{cases} \beta_0 \exp\left(-\frac{\delta^2_{r,ij}}{2\,\text{receptive field size}^2}\right) & \text{for } i \neq j \\ 0 & \text{for } i = j \end{cases} \tag{1}$$

where β_0 is the standard inhibition parameter and $\delta_{r,ij}$ is the Euclidean distance between the receptive fields of nodes i and j.

Dynamics

The dynamics of the network are governed by the following ordinary differential equation (following[1,2,11]):

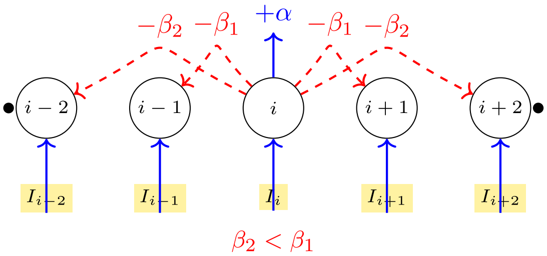


Fig. 1. Illustration of a sparse recurrent neural network with five nodes in a linear array. All the nodes exhibits self-excitation, represented by $+\alpha$ above the central node, while lateral inhibition between neighboring nodes is given by β_1 and inhibition between nodes separated by another is shown as β_2($\beta_2 < \beta_1$). Transient input is shown as I_i below each node.

$$\frac{dx_i}{dt} = -x_i + \alpha F(x_i) - \sum_{j=1, j\neq i}^{N} \beta_{ij} F(x_j) + I_i + \text{noise} \tag{2}$$

where $F(x)$ is defined as:

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{x}{1+x} & \text{for } x > 0 \end{cases} \tag{3}$$

Here, x_i represents the state of the i-th node, α is a scaling factor, I_i is the input to node i, and noise is drawn from a Gaussian distribution with mean zero and standard deviation of 0.03.

The transient input is applied to a subset of nodes for a duration of $T = 1$ second, after which the network dynamics evolve based solely on the interactions among nodes.

Non-divergence condition

At steady state, we have $\frac{dx_i}{dt} = 0$. Therefore, Eq. (2) simplifies to:

$$x[n] = \alpha F(x[n]) - \sum_{j=1, j\neq i}^{N} \beta_{ij} F(x[n]). \tag{4}$$

This simplifies further to:

$$x[n] = F(x[n]) \left(\alpha - \sum_{j=1, j\neq i}^{N} \beta_{ij} \right). \tag{5}$$

To address the non-divergence condition using the perturbation method, we can start by considering small perturbations around a steady-state solution. Let's denote the steady-state solution as x_i^0 and the perturbation as ϵ_i, such that $x_i = x_i^0 + \epsilon_i$.

Given the differential equation:

$$\frac{dx_i}{dt} = -x_i + \alpha F(x_i) - \sum_{j=1, j\neq i}^{N} \beta_{ij} F(x_j) + I_i + \text{noise} \tag{6}$$

we can linearize it around the steady-state by substituting $x_i = x_i^0 + \epsilon_i$ and expanding $F(x_i)$ using a Taylor series:

$$F(x_i) \approx F(x_i^0) + F'(x_i^0)\epsilon_i \tag{7}$$

Substituting this into the original equation and neglecting higher-order terms, we get:

$$\frac{d\epsilon_i}{dt} = -\epsilon_i + \alpha F'(x_i^0)\epsilon_i - \sum_{j=1, j\neq i}^{N} \beta_{ij} F'(x_j^0)\epsilon_j \tag{8}$$

For the non-divergence condition, we require that the perturbations do not grow unbounded over time. This implies that the eigenvalues of the linearized system must have non-positive real parts.

The linearized system can be written in matrix form as:

$$\frac{d\boldsymbol{\epsilon}}{dt} = \mathbf{A}\boldsymbol{\epsilon} \tag{9}$$

where $\boldsymbol{\epsilon}$ is the vector of perturbations and $\mathbf{A}$ is the Jacobian matrix of the system evaluated at the steady-state. The elements of $\mathbf{A}$ are given by:

$$\begin{aligned} A_{ii} &= -1 + \alpha F'(x_i^0) \\ A_{ij} &= -\beta_{ij} F'(x_j^0) \quad \text{for } i \neq j \end{aligned} \tag{10}$$

To ensure non-divergence, we need to analyze the eigenvalues of $\mathbf{A}$. If all eigenvalues have non-positive real parts, the system is stable and perturbations will not grow unbounded.

For stability, we require that the real parts of all eigenvalues λ of the matrix $\mathbf{A}$ satisfy $\mathrm{Re}(\lambda) \leq 0$.

The eigenvalues λ of $\mathbf{A}$ are found by solving the characteristic equation $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$.

To bound the eigenvalues, we apply the Gershgorin Circle Theorem. According to the theorem, each eigenvalue λ lies within at least one of the Gershgorin disks, centered at the diagonal element A_{ii} with radius equal to the sum of the absolute values of the off-diagonal elements in row i:

$$|\lambda - A_{ii}| \leq \sum_{j \neq i} |A_{ij}| \tag{11}$$

The diagonal terms are:

$$A_{ii} = -1 + \alpha F'(x_i^0) \tag{12}$$

For stability, we require $A_{ii} \leq 0$, which gives the condition:

$$-1 + \alpha F'(x_i^0) \leq 0 \tag{13}$$

This implies:

$$\alpha \leq \frac{1}{F'(x_i^0)} \tag{14}$$

The off-diagonal terms are:

$$A_{ij} = -\beta_{ij} F'(x_j^0) \tag{15}$$

The sum of the absolute values of the off-diagonal terms in row i is:

$$\sum_{j \neq i} |A_{ij}| = \sum_{j \neq i} \beta_{ij} F'(x_j^0) \tag{16}$$

Using the Gershgorin Circle Theorem, we require that the maximum eigenvalue $\lambda_{\max}$ satisfies:

$$|\lambda_{\max} - A_{ii}| \leq \sum_{j \neq i} |A_{ij}| \tag{17}$$

For stability, $\lambda_{\max} \leq 0$. Substituting $A_{ii} = -1 + \alpha F'(x_i^0)$, we get:

$$1 - \alpha F'(x_i^0) \geq \sum_{j \neq i} \beta_{ij} F'(x_j^0) \tag{18}$$

To ensure non-divergence, we require:

$$\alpha F'(x_i^0) \leq 1 - \sum_{j \neq i} \beta_{ij} F'(x_j^0) \tag{19}$$

This gives an upper bound on α. For a conservative estimate, we can approximate the condition by assuming the right-hand side is maximized over all nodes i:

$$\alpha \leq \frac{1}{\max_{i,j} \left(\sum_{j \neq i} \beta_{ij} F'(x_j^0) \right)} \tag{20}$$

This condition ensures that the largest eigenvalue $\lambda_{\max} \leq 0$, guaranteeing stability and non-divergence.

Energy function and stability criterion

A further stability analysis can be conducted based on an energy function formalism given by[17]. Following that, we define the node-wise energy function as:

$$dH_i \propto \dot{x}_i d\dot{x}_i, \tag{21}$$

and the total energy of the network is given by:

$$H = \sum_{i=1}^{N} \int dH_i. \tag{22}$$

The Lyapunov stability criterion states that the system is stable if:

$$H \leq 0. \tag{23}$$

We perturb the system around its steady-state solution x_i^0 by introducing small perturbations ϵ_i, such that:

$$x_i = x_i^0 + \epsilon_i \tag{24}$$

Substituting this into the system's governing differential equation:

$$\frac{dx_i}{dt} = -x_i + \alpha F(x_i) - \sum_{j \neq i} \beta_{ij} F(x_j) + I_i + \text{noise} \tag{25}$$

and expanding the activation function $F(x)$ using a Taylor series around x_i^0, we get:

$$F(x_i) \approx F(x_i^0) + F'(x_i^0)\epsilon_i \tag{26}$$

Neglecting higher-order terms, the linearized equation for ϵ_i becomes:

$$\frac{d\epsilon_i}{dt} = -\epsilon_i + \alpha F'(x_i^0)\epsilon_i - \sum_{j \neq i} \beta_{ij} F'(x_j^0)\epsilon_j \tag{27}$$

We now express the energy change dH_i in terms of the perturbations ϵ_i and their derivatives. Given $dH_i \propto \dot{x}_i d\dot{x}_i$, and noting that $\dot{x}_i = \frac{d\epsilon_i}{dt}$, the energy for each node is:

$$dH_i = \frac{1}{2}\dot{\epsilon}_i^2 \tag{28}$$

Thus, the total energy of the system becomes:

$$H = \sum_{i=1}^{N} \frac{1}{2}\dot{\epsilon}_i^2 \tag{29}$$

For the system to be stable, we require that the total energy H is non-increasing, i.e., $H \leq 0$. To derive the stability condition, we take the time derivative of the energy function:

$$\frac{dH}{dt} = \sum_{i=1}^{N} \dot{\epsilon}_i \ddot{\epsilon}_i \tag{30}$$

Substituting the linearized equation for $\ddot{\epsilon}_i$:

$$\ddot{\epsilon}_i = \frac{d^2\epsilon_i}{dt^2} = -\epsilon_i + \alpha F'(x_i^0)\epsilon_i - \sum_{j \neq i} \beta_{ij} F'(x_j^0)\epsilon_j \tag{31}$$

we obtain:

$$\frac{dH}{dt} = \sum_{i=1}^{N} \dot{\epsilon}_i \left(-\epsilon_i + \alpha F'(x_i^0)\epsilon_i - \sum_{j \neq i} \beta_{ij} F'(x_j^0)\epsilon_j \right) \tag{32}$$

For stability, we require $\frac{dH}{dt} \leq 0$, which leads to the condition:

$$\sum_{i=1}^{N} \dot{\epsilon}_i \left(-\epsilon_i + \alpha F'(x_i^0)\epsilon_i - \sum_{j \neq i} \beta_{ij} F'(x_j^0)\epsilon_j \right) \leq 0 \tag{33}$$

By focusing on the dominant diagonal elements (self-interaction terms), we can approximate the condition as:

$$\alpha \leq \frac{1}{F'(x_i^0)} \tag{34}$$

Thus, the stability condition can be written as:

$$\alpha \leq \frac{1}{\max_{i,j}\left(\sum_{j \neq i} \beta_{ij} F'(x_j^0)\right)} \tag{35}$$

This is the same stability criterion derived from the perturbation method.

Simulating single layer of distance dependent recurrent OCOS

We simulated a sparse recurrent on-center off-surround neural network consisting of $N = 64$ nodes arranged in a square grid topology. Each node interacts with its neighbors through self-excitation and distance-based lateral inhibition. The dynamics of each node are governed by the Eqs. (2), (3) and (1). Some parameter choices for the simulation (e.g. noise, α, etc) were inspired by convergence tests performed on fully connected versions of the additive recurrent OCOS network[2,11].

Simulation procedure: pseudocode

The following pseudocode outlines the main steps in simulating the network dynamics using Euler's method for numerical integration:

```
**Input:** $N, \alpha, \beta_0, \sigma, dt, T_{end}, input\_duration, num\_inputs$
**Output:** $x\_history$
Initialize $x \leftarrow \mathbf{0}$ {Initial node activity}
Compute $node\_positions$ for 2D grid of size $\sqrt{N} \times \sqrt{N}$
Compute $beta\_matrix$ using distance-based inhibition rule
Generate random input indices for $num\_inputs$ nodes
Set $input \leftarrow 1$ for the selected nodes
**for** $t \leftarrow 1$ to $T_{end}/dt$ **do**
    **if** $t \leq input\_duration/dt$ **then**
        $I \leftarrow$ input to selected nodes
    **else**
        $I \leftarrow \mathbf{0}$ {No external input after $T = 1$}
    **end if**
    Compute $dxdt \leftarrow -x + \alpha F(x) - \beta_{ij} \sum_{j \neq i} F(x_j) + I + \text{noise}$
    Update $x \leftarrow x + dxdt \cdot dt$
    Store $x$ in $x\_history$
**end for**

**return** $x\_history$
```

Algorithm 1. Additive on-center off-surround recurrent neural network.

The model simulates a sparse, additive recurrent neural network where each node represents a region of space in the visual field. The main characteristics of the model are:

- The *spatial distribution* of objects (either close together or far apart) affects the memory capacity of the network.
- The size of the *receptive field* for each node controls how much inhibition from neighboring nodes affects the node's activity. Smaller receptive fields yield higher spatial precision, while larger receptive fields blend inputs over a wider area.
- The *error rate* (measured as the Hamming distance between input and output patterns) is used as a metric to quantify how well the network can store and retrieve spatial information.
- Inhibition between nodes is *distance-dependent*, decaying exponentially with the Euclidean distance between them.

Simulation framework

The simulation runs over multiple time steps, with noise added to the system to simulate biological variability. Each simulation is repeated multiple times to calculate average performance.

The receptive field size significantly influences how nodes interact with their neighbors. Small receptive fields enable high spatial resolution, allowing for precise localization of stimuli. In contrast, larger receptive fields average input over a broader area, which can enhance capacity but may reduce the precision of object representation.

Two types of spatial distribution are simulated:

- Close inputs: Stimuli are located around the central nodes, defined as being clustered together in a small region.
- Far inputs: Stimuli are placed radially 3 nodes away from the center, resulting in a widely separated distribution across the grid.

Table 1 summarizes the key parameters used in the simulation:

Parameter	Description	Value
N	Number of nodes	64 (8 × 8 grid)
dt	Time step	0.01
T_{total}	Total simulation time	10
T_{input}	Input duration	1
σ_{noise}	Noise standard deviation	0.03
α	Activation scaling factor	2
β_0	Base inhibition strength	0.2-0.3
r_{low}	Low receptive field size	0.5
r_{high}	High receptive field size	32
num_simulations	Number of simulations per condition	100
set_sizes	Number of objects (stimuli)	1 to 12

Table 1. Key parameters for the simulation.

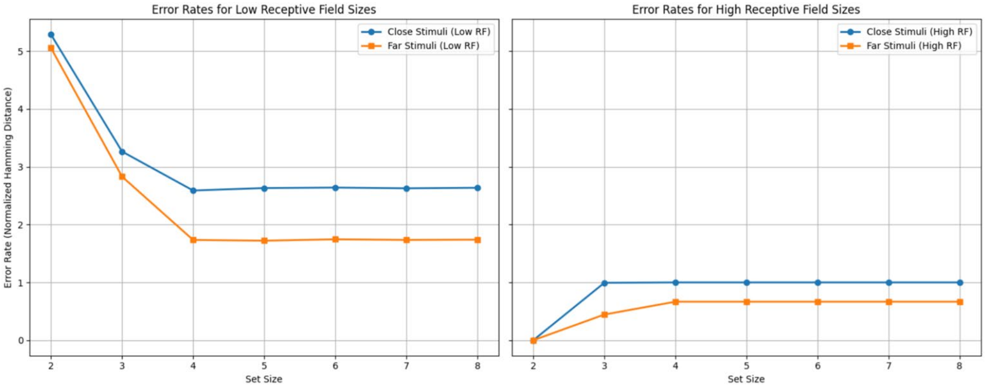


Fig. 2. Error rates (as measured by Hamming distance) for different receptive field sizes when processing close and far-apart stimuli in the network ($\alpha = 2.0$, $\beta = 0.3$).

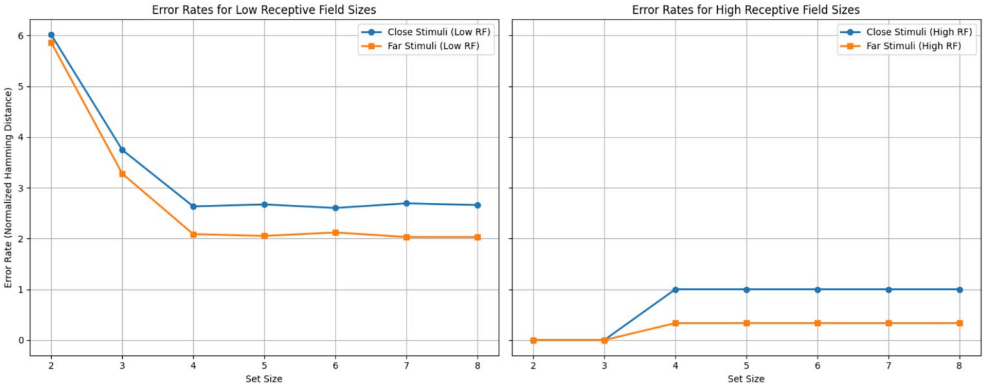


Fig. 3. Error rates (as measured by Hamming distance) for different receptive field sizes when processing close and far-apart stimuli in the network ($\alpha = 2.0$, $\beta = 0.2$).

Simulation results

The results presented in Figs. 2 and 3 represent the error rates (as measured by Hamming distance) for different receptive field sizes when processing close and far-apart stimuli in the network. The figures compare the network's performance under two distinct inhibitory regimes ($\beta = 0.3$ vs $\beta = 0.2$) while maintaining a constant excitation parameter ($\alpha = 2.0$).

For low receptive field sizes (averaged for sizes 0.5, 1, and 2) given in left panels of Figs. 2 and 3

- *Close stimuli* The error rate starts high (around 5.5 Hamming distance) when the set size is 2, indicating that the network struggles to accurately encode or differentiate the stimuli when the receptive field size is small. As

the set size increases to 4, the error rate drops sharply, stabilizing at around 2.5, meaning the network becomes more consistent in its performance. This suggests that for small receptive fields, the network can resolve more tightly packed inputs as more objects are added, up to a certain point (around set size 4).
- *Far stimuli* The network performs better with far stimuli, with error rates starting lower (around 2.5) for setsize 3 and dropping quickly to around 1.5 for a set size of 4, before stabilizing. This indicates that small receptive fields are advantageous for spatially dispersed inputs, allowing for better resolution.

For high receptive field sizes (averaged for sizes 4, 8, and 16) given in right panels of Figs. 2 and 3

- *Close stimuli* The error rate starts low but increases sharply to about 1 as the set size grows to 4, where it then plateaus. This pattern indicates that with larger receptive fields, the network struggles to accurately differentiate close stimuli as more objects are introduced. The larger fields likely result in overlap between neural representations, leading to less precision in distinguishing between objects.
- *Far stimuli* Similar to the close stimuli, the far stimuli show increasing error rates as set size increases. However, the error rates for far stimuli stay consistently lower than those for close stimuli, even with larger receptive fields. This implies that large receptive fields can still handle spatially distributed stimuli relatively well, but not as effectively as small receptive fields.

These figures demonstrate that a higher inhibitory strength leads to lower error rates for smaller setsizes for low receptive field sizes. On the other hand for high receptive field sizes lower inhibition leads to lower error rates, as evidenced by the reduced Hamming distance across various receptive field sizes and stimulus distributions. Specifically, stronger inhibition ($\beta = 0.3$) appears to more effectively counteract the excitatory spread, reducing overlap between neural representations and thus minimizing encoding errors for low receptive field sizes. This is particularly pronounced for far-apart stimuli. In contrast, lower inhibition ($\beta = 0.2$) is associated with higher error rates for low receptive field sizes, but does much better at higher receptive field sizes—highlighting the importance of balanced excitation–inhibition dynamics in our model.

In our simulations, the selection of nodes for stimulation is based on the spatial distribution that reflects the receptive field (RF) properties of the neurons. Specifically, for "close inputs" we uniformly sample nodes from a small, predefined cluster within the 8×8 grid. In our implementation, these nodes are chosen uniformly at random from a restricted region (e.g., the top-left 2×2 sub-grid), which is intended to represent a localized group of neurons with small RFs. For "far inputs", nodes are selected from predetermined, spatially distant locations, ensuring that stimuli are presented at the periphery of the grid. Although the external input I is applied in a binary (0/1) fashion - i.e., a value of 1 is delivered to the selected nodes and 0 elsewhere—the subsequent neural response in each node is further modulated by a Gaussian tuning function $F(x)$. This Gaussian function models the intrinsic receptive field profile of the neuron and is separate from the uniform (binary) input. Thus, while the input is uniform in its application, the effective activation pattern reflects the spatial weighting imposed by the neurons' receptive fields. This dual approach allows us to capture both the uniform external stimulation and the biologically realistic Gaussian modulation due to RF size.

The differences in error rates observed for low versus high receptive field sizes can be understood by considering the encoding errors that result from overlapping receptive fields within a single layer. In our model, each neuron's receptive field is characterized by a Gaussian profile,

$$R(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right),$$

and the overlap between two neurons, whose receptive fields are centered a distance d apart, is given by

$$O(d) = \exp\left(-\frac{d^2}{4\sigma^2}\right).$$

When the network receives external input, ideally, the neurons would respond in a manner that minimizes ambiguity. However, if only a fraction $A(L)$ of the available input is effectively processed (due, for example, to limited feedback or incomplete recurrent integration), then the effective input is degraded by the residual interference arising from overlapping receptive fields. We propose that the effective clarity of the encoded stimulus is attenuated by a cross-talk factor that captures this interference, which we define as

$$C(L) = A(L) \exp\left[-B\left(1 - A(L)\right)^2\right],$$

where B is a scaling constant and $A(L)$(ranging from 0 to 1) reflects the fraction of effective processing. When $A(L) = 1$(full effective processing), $C(L) = 1$ and the stimulus is encoded unambiguously. Conversely, if $A(L) < 1$, the resulting overlap errors, which scale approximately with $(1 - A(L))^2$, reduce $C(L)$, thereby leading to higher error rates. Although high receptive field sizes (e.g., corresponding to intermediate values such as 4, 8, and 16) yield a lower overall Hamming distance, indicating robust global performance, increased receptive field size also promotes overlap. Such overlap can deteriorate spatial precision, particularly for closely spaced stimuli. A complete characterization of network performance must therefore consider both the global error and the cross-talk-induced encoding errors.

The results show that small receptive fields provide better precision for far-apart stimuli, while larger receptive fields reduce precision for close stimuli due to overlap. However, smaller receptive fields perform poorly when there are fewer objects, particularly for close stimuli, because the network may not effectively capture the finer details at low set sizes.

There is a clear trade-off between capacity and precision in the network. Small receptive fields improve spatial precision, especially for far-apart stimuli, but reduce capacity. Larger receptive fields improve capacity by pooling information over a larger area but at the cost of spatial precision, especially when stimuli are close together.

Modeling visual working memory

We propose that VWM is structured as a hierarchical processing network operating in parallel and physically yoked to the visual processing network. Visual processing from lower visual areas to higher cortical areas is a combination of a feed-forward, stimulus-driven process and top-down recurrent modulation. The information encoded through this process may be passed through to the memory network operating alongside it.

A feed-forward pass through the Visual Hierarchy (VH, it encompasses brain structures from lower visual areas such as LGN and V1 to higher brain regions like IT involved in direct visual processing[18]) typically takes approximately 100–150 ms[13], at which point information may begin to impact a person's memory and behavior[19]. However, the information available at this point is relatively coarse (such as scene category, rather than knowledge of specific scene details), and more detailed knowledge requires further processing time[20]. Given that visual information has passed fully through the VH by the end of the first 150 ms, this additional refinement of information must necessarily come from a combination of additional fixations and attentional processes[18].

VWM hierarchy runs parallel to VH, with the top layer being in prefrontal cortex (PFC). The best physiological candidate to house the top layer of VWM is in the PFC, which has sufficient connectivity to higher visual areas[15,21] and for integrating executive control signals with working memory operation. The abstract PFC layer consists of nodes that anchor a memory sample each. The other layers of VWM are yoked to the the layers of VH. We have named them R_{V1} to R_{IT} for convenience. During a feed forward pass visual information is processed by the VH until selection takes place in IT. Following the selection in IT, a control signal from the executive system activates the PFC node corresponding to the current abstract memory sample. As VH neurons have their own receptive fields that get bigger as one moves up the VH, the selected IT neurons can “see” only the few neurons at the lower layer (say TEO) that the IT neuron receives input from. Thus the memory content associated with the abstract node at PFC for the current visual input during the feed-forward process is just the IT neuron and the gating information for that particular neuron. The gating neurons that control the flow of information through the VH have both feature and location information. So the memory content is the selected IT neuron and the gating structure connected to it.

Following the feed-forward pass if the visual system needs more detailed information to complete the task at hand, the feedback pass moves down the VH to complete the selection process. Thus the memory content is also refined during the same process, and subsequent layers of VWM are given active input to relevant nodes and connected to the abstract memory sample node at PFC. Active input to VWM stops with top-down disengagement signal that can originate with new stimulus, gaze shift, blink, or covert attentional shift. Following disengagement, the VWM hierarchy tries to maintain the memory sample.

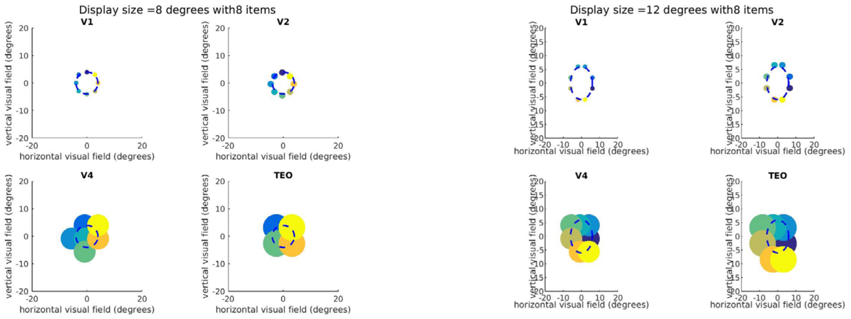
In the following, we apply a hierarchical recurrent neural network (HRNN) model to simulate the dynamics of VWM. In this architecture, the abstract or top layer of VWM, theorized to reside in the prefrontal cortex (PFC), represents a single memory sample and is connected to the lower layers of the VWM hierarchy through simple, yet effective, rules. The non-abstract layers of the VWM hierarchy contain copies of gating neurons, positioned between successive layers of the visual hierarchy. These gating neurons encode both feature-based and location-based information. As a result, the non-abstract layers can be formally described as containing nodes that have conjunctive properties of both feature and location. This paves the way for us to model feature-based center-surround activity in VWM in future models, as some authors have suggested[22].

Model description

The proposed model includes the following key components:

- Input layer: The input layer serves as the initial stage where objects first appear as single nodes, capturing raw object information. This layer is fully connected and can incorporate salient or peripheral items from the dorsal visual pathway.
- Feature layers: Following the input layer, each subsequent layer represents distinct feature dimensions, such as location, orientation, and color. Nodes within each feature layer have tuning curves specific to their feature and are organized topographically according to their receptive field centers.
- Topographic representation: Each feature layer is topographically organized, with spatially distributed nodes sensitive to specific locations and features within their receptive fields.
- Attention and information clarity: The clarity of feature representations in each layer depends on attentional individuation. Successful attentional focusing enables clearer feature representation in the corresponding layer, while unattended features are represented less distinctly.
- Recurrent dynamics: The model incorporates recurrent OCOS dynamics within and between layers. Each layer has local excitatory (center) and broader inhibitory (surround) connections, with distance-dependent inhibition governing suppression based on spatial proximity.

In a spatiotopic hierarchical network with branching connections, there will necessarily be overlapping connections in upper layers between sufficiently nearby yet distinct loci of activity in the input layer (this is an issue commonly referred to as cross-talk). As one moves further up the layers of the network (where receptive fields become progressively larger), this issue of overlap gets progressively worse (See Fig. 4).



(a) For a standard 10 x 10 degree display of objects (roughly 1 degrees in diameter) even if they are separated maximally from each other along the circumference of a circle, the large receptive field sizes in higher visual areas constrain the available processing capacity to roughly 4 objects.

(b) If we increase the display size slightly it allows for more object nodes to receive information in the higher visual areas, leading to the possibility that more items will be remembered.

Fig. 4. Receptive field overlaps due to larger receptive fields at higher visual areas and cross-talk between the feed-forward processing paths.

The specificity of features represented in the memory layers depends on the depth of feedback processing achieved within the VH, modulated by a cross-talk function, $C(L)$, which controls the level of feature clarity available to the memory network (Fig. 5).

To capture the partially localized representation in the Visual Hierarchy (VH) as a p-Lattice, we model the cross-talk function $C(L)$ to reflect the evolving clarity and interconnectedness of features in the memory network as feedback processing deepens.

The p-lattice structure and feedback levels

The p-Lattice represents a partially ordered set of localized features, where each node in the lattice corresponds to a feature with spatial and/or feature specificity, and the highest (root) node corresponds to the object level. A fully processed recurrent representation in VH corresponds to the entire lattice being visible and fully ordered, reflecting complete feature specificity across levels. As feedback processing deepens, only a subset of this lattice becomes accessible, introducing uncertainty.

Cross-talk function $C(L)$ with p-lattice representation

We introduce a cross-talk function, $C(L)$, that is modulated by the fraction of the p-Lattice that has been processed at each feedback level L. At lower levels, only a small fraction of the lattice structure is visible, introducing uncertainty due to partial feature information. Let:

- $\mathcal{L}$ represent the p-Lattice.
- $|\mathcal{L}|$ denote the total number of nodes in the lattice.
- $|\mathcal{L}_L|$ represent the subset of nodes visible at feedback level L.

Define $A(L)$ as the fraction of the lattice that is visible at feedback level L:

$$A(L) = \frac{|\mathcal{L}_L|}{|\mathcal{L}|} \tag{36}$$

This fraction, $A(L)$, serves as a metric for feature clarity. It indicates the completeness of the feature information reaching the VWM, where $A(L) = 1$ implies full lattice visibility (complete recurrent processing to V1) and $A(L) < 1$ implies partial lattice visibility.

The cross-talk function $C(L)$ can now be constructed to reflect this hierarchical and progressive access to feature information. We propose that $C(L)$ depends on $A(L)$ as follows:

$$C(L) = A(L) \cdot \exp(-B(1 - A(L))^2) \tag{37}$$

where B is a scaling constant that controls the rate at which cross-talk (feature clarity) increases with feedback processing depth.

- At $L = 1$(feed-forward only), $A(L)$ is low, so $C(L) \approx 0$, reflecting high uncertainty due to minimal feature information.
- At $L = 3$(full feedback), $A(L) = 1$, so $C(L) = 1$, reflecting full feature clarity as the entire lattice is visible.

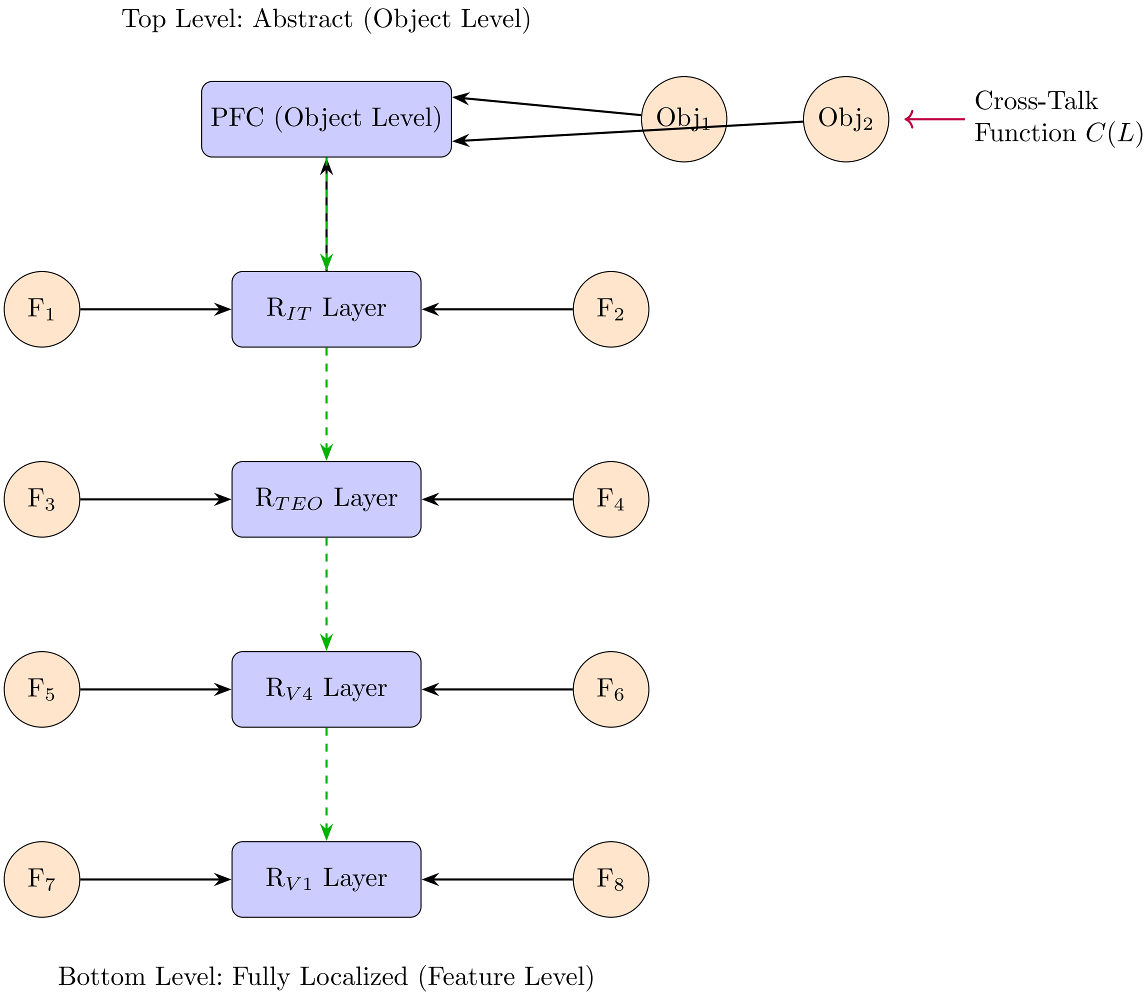


Fig. 5. Hierarchical structure of the VWM model alongside the Visual Hierarchy (VH). The network layers represent different stages of feature processing, from abstract object level at the prefrontal cortex (PFC) down to localized feature processing in V1. The cross-talk function $C(L)$ adjusts feature clarity based on feedback processing depth, from object identification at higher layers to specific feature localization at lower layers. Dashed arrows indicate feedback loops.

The incomplete visibility of the p-Lattice can be seen as an uncertainty factor in the feature layers of VWM. We define an uncertainty metric $U(L)$ as:

$$U(L) = 1 - A(L) \tag{38}$$

Thus, the cross-talk function $C(L)$ implicitly incorporates uncertainty, as lower values of $A(L)$ correspond to higher uncertainty in the feature representation.

Incorporating $C(L)$ into the dynamic equation for the activation x_i of neuron i in the feature layer, we get:

$$\frac{dx_i}{dt} = -x_i + C(L) \cdot \left(\sum_{j \in \text{neighbors}} W_{ij} \cdot R_j(f) - \lambda \sum_{k \in \text{network}} G_{ik} \cdot x_k \right) \tag{39}$$

where:

- x_i: Activation of neuron i in the feature layer,
- $C(L)$: Cross-talk function modulating information clarity based on VH feedback depth L,
- W_{ij}: Excitatory connections between neuron i and neighboring neurons j,
- $R_j(f)$: Gaussian-tuned response of neuron j to feature f,
- λ: Scaling factor for inhibition,

- $G_{ik} = \exp\left(-\frac{d_{ik}^2}{2\sigma_d^2}\right)$: Distance-dependent inhibition, where d_{ik} is the distance between neurons i and k, and σ_d controls the spatial extent of inhibition.

The Gaussian tuning for feature f can be represented as:

$$R_j(f) = \exp\left(-\frac{(f-f_j)^2}{2\sigma_f^2}\right) \tag{40}$$

where f_j is the preferred feature value for neuron j, and σ_f controls the feature selectivity of the neuron.

With this construction, the model operates as follows:

- $L = 1$: $C(L) \approx 0$, high uncertainty, and only coarse object-level information reaches the memory layers.
- $L = 2$: $C(L)$ increases, localization and complex features (like orientation and size) begin to be integrated as more of the p-Lattice is visible.
- $L = 3$: $C(L) \approx 1$, full clarity, and all feature specifics (precise orientation, texture, color) are encoded in memory as the entire lattice becomes visible.

Derivation of the cross-talk function

Consider a single layer of a sparse recurrent on-center off-surround (OCOS) network with N nodes arranged in a 2D grid. Each node has a receptive field characterized by a Gaussian function,

$$R(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right), \tag{41}$$

where σ represents the effective receptive field size. When two nodes with centers separated by a distance d interact, the degree of overlap between their receptive fields is given by

$$O(d) = \int_{-\infty}^{\infty} R(x)R(x-d)\,dx = \exp\left(-\frac{d^2}{4\sigma^2}\right), \tag{42}$$

up to a normalization constant.

In an ideal, fully localized representation—where recurrent processing has engaged the entire p-lattice—the effective stimulus is represented unambiguously and thus the cross-talk modulation is maximum, i.e. $C(L) = 1$ when the fractional coverage $A(L) = 1$. However, when only a fraction $A(L) < 1$ of the p-lattice is active (reflecting limited feedback), not only is the number of neurons contributing reduced, but the ambiguity introduced by overlapping receptive fields (quantified by $O(d)$) also degrades the effective encoding

Assuming that the average effective overlap among the activated nodes is $\bar{O}$, the loss in clarity due to overlapping is proportional to the fraction of the p-lattice that is missing, i.e. $(1 - A(L))$. If the encoding error due to overlap scales as the square of this missing fraction, we can write the corrective factor as

$$\phi\big(1 - A(L)\big) = \exp\left[-B\,\left(1 - \bar{O}\,A(L)\right)^2\right], \tag{43}$$

where B is a positive scaling constant and $\bar{O}$ represents the effective overlap when all nodes are active. In many cases, when the full p-lattice is engaged, we expect $\bar{O}$ to approach 1, recovering full clarity.

Thus, the effective cross-talk function can be expressed as the product of two factors: (i) the linear scaling due to the fraction of the p-lattice that is activated, and (ii) an exponential penalty representing the encoding error from incomplete overlap resolution:

$$C(L) = A(L)\cdot\exp\left[-B\,\left(1 - \bar{O}\,A(L)\right)^2\right]. \tag{44}$$

When $A(L) = 1$, corresponding to full feedback, Eq. (44) yields $C(L) = 1($ assuming $\bar{O} = 1)$. Conversely, when $A(L) < 1$, the cross-talk function $C(L)$ is reduced due to the compounded effect of fewer active neurons and increased ambiguity from receptive field overlap. This formulation encapsulates how the clarity of feature encoding in the visual working memory network improves with deeper recurrent localization, thereby reducing the encoding error induced by overlapping receptive fields.

For low receptive field sizes (e.g., r_{low} = 0.5, 1, 2), the network excels at resolving far-apart stimuli, with error rates decreasing as the set size increases up to a plateau. This precision arises from limited inhibition among nodes, preserving spatial detail. However, for close stimuli, small receptive fields struggle at smaller set sizes, reflecting difficulty in resolving tightly packed inputs due to insufficient pooling of activity.

Conversely, high receptive field sizes (e.g., r_{high} = 4, 8, 16) demonstrate the network's ability to pool over larger areas, increasing capacity. However, this comes at the cost of *reduced spatial precision*, particularly for close stimuli, where overlap between neural activations causes an increase in error rates. For far-apart stimuli, larger receptive fields still perform reasonably well, as the spatial separation minimizes overlap.

The distance-dependent inhibition, combined with noise, introduces variability that mimics biological systems. Figures 2 and 3 reveal the sensitivity of performance to β, with stronger inhibition (β = 0.3) generally yielding better discrimination than weaker inhibition ($\beta = 0.2$). Overall, the network demonstrates a clear trade-off between precision and capacity, depending on receptive field size and stimulus distribution. These findings align with the theoretical framework of spatially distributed memory systems.

In this way, the cross-talk function $C(L)$ allows the VWM model to adaptively refine its feature representation based on feedback processing depth, encapsulating the progression from coarse to detailed visual memory as the lattice visibility increases. This captures the gradual resolution of feature information in working memory, dependent on recurrent localization and attentional processing within the VH.

Modeling simple hierarchical VWM

Simulation parameters and methodology

The hierarchical VWM model was simulated using a sparse additive recurrent neural network (RNN) incorporating distance-dependent inhibition and a cross-talk function to mimic a p-lattice hierarchy (Scene $\rightarrow$ Objects $\rightarrow$ Features). The model features two distinct feature layers (orientation and size) whose responses are modulated by top-down feedback. In our simulation, a fraction $A(L)$ of the p-lattice is visible depending on the feedback level L, with the cross-talk function defined as

$$C(L) = A(L)\exp\left(-B(1 - A(L))^2\right).$$

For instance, we set $A(1) = 0.3$, $A(2) = 0.7$, and $A(3) = 1.0$, corresponding to progressively refined representations.

Key simulation parameters are summarized in Table 2. In each trial, a subset of objects (set sizes 2, 4, or 8) from a pool of 20 objects is encoded into the feature layers. Object features (orientation and size) are generated using Gaussian tuning curves, while distance-dependent inhibition (with a Gaussian decay controlled by σ_d) governs local interactions. During the encoding phase (first 30% of the simulation time), feature inputs are added to the network, scaled by the cross-talk clarity $c_{\text{talk}} = C(L)$. The recurrent dynamics then allow the network to maintain these features during the delay period. Recall is assessed by comparing the final states of the orientation and size layers with ideal reference patterns.

Algorithm overview

The simulation algorithm proceeds as follows:

Parameter	Description	Value
N	Number of nodes (grid size)	64 (8×8)
dt	Time step	0.01 s
T_{total}	Total simulation time	10 s
T_{input}	Encoding duration	30% of T_{total}
$noise\_std$	Noise standard deviation	0.02
α	Self-excitation	2.0
β_0	Base lateral inhibition strength	0.1
σ_d	Inhibition decay (receptive field size)	2.0
B	Cross-talk scaling constant	5.0
Set Sizes	Number of objects	2, 4, 8
Num. Runs	Repetitions per condition	100

Table 2. Simulation parameters for the hierarchical VWM model.

```
Input: set size, total objects, simulation parameters.
Build distance-dependent inhibition matrix.
Initialize feature layer states for orientation and size.
Generate object features using Gaussian tuning.
Determine feedback level $L$ and compute $C(L)$.
for each timestep $t$ do
    if $t < 0.3 \times T_{\text{total}}$ then
        for each object $o$ do
            Add cross-talk scaled Gaussian input to feature layers.
        end for
    end if
    Update feature states via recurrent dynamics:
        $x(t+1) = x(t) + dt\Big[-x(t) + \alpha F(x(t)) - \beta_{\text{mat}} F(x(t)) + \text{noise}\Big]$
end for
Compute recall probability by measuring overlap between final states and reference patterns.
```

Algorithm 2. Hierarchical VWM Simulation

Visualization of the hierarchical p-lattice structure

Figure 6 schematically illustrates the hierarchical structure of our VWM model. The top layer (Scene) projects onto object nodes, which in turn propagate feature-specific information (Orientation and Size) through local, distance-dependent recurrent dynamics. The cross-talk function $C(L)$(displayed on the right) modulates feature clarity based on the feedback level.

Recall probability

Figure 7 displays the recall probability as a function of set size (2, 4, and 8), averaged over 100 simulation runs with error bars representing the standard error of the mean. Our results demonstrate a decline in recall probability with increased set size, consistent with behavioral findings that working memory capacity is limited. Notably, the model reveals that when the p-lattice is fully visible (i.e., $L = 3$ and $C(L) = 1$), recall performance is maximized. In contrast, lower cross-talk values associated with partial feedback ($L = 1$ or $L = 2$) lead to degraded recall performance, underscoring the role of top-down feedback in sharpening feature representations and enhancing memory capacity.

Our hierarchical VWM model successfully integrates top-down feedback, distance-dependent inhibition, and a p-lattice cross-talk mechanism to simulate multi-item working memory.

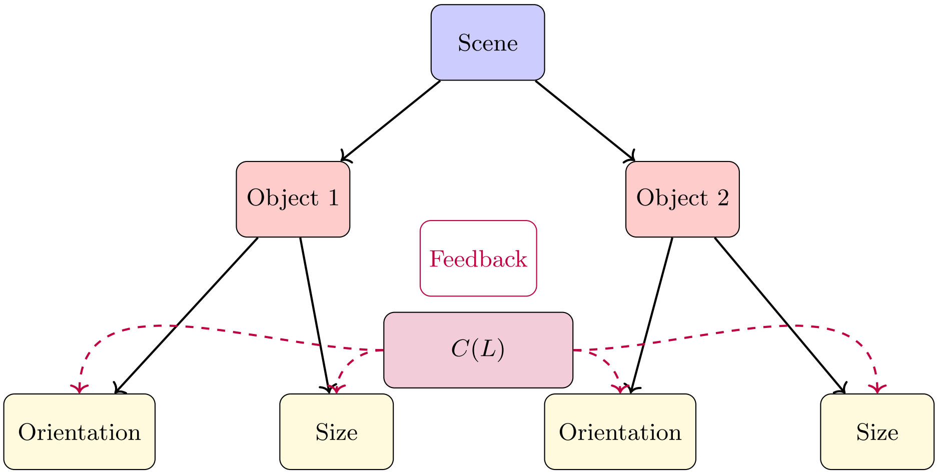


Fig. 6. Schematic representation of the hierarchical visual working memory (VWM) model. The scene is parsed into individual Object representations, each of which is further resolved into feature dimensions (orientation and size). the cross-talk function $C(L)$ provides feedback to the feature level, dynamically modulating precision based on the degree of overlap and load.

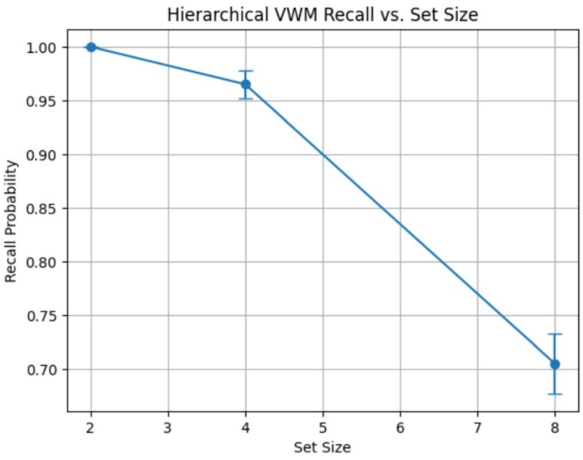


Fig. 7. Recall probability versus set size for the hierarchical VWM model. Data points represent averages over 100 simulation runs with error bars (SEM). The results highlight decreased recall probability with increased set size, demonstrating the capacity limitations intrinsic to the model.

Discussion

In the current work we started with simulating the dynamics of a recurrent on-center off-surround neural network layer with distance dependent inhibition implementation (see Fig. 1). The simulation results highlight the trade-offs between capacity and precision in the network, as demonstrated in Figures 2 and 3 and summarized in Table 1. The error rates (measured via Hamming distance) provide key insights into how receptive field size, spatial distribution of stimuli, and inhibition strength (β) impact the network's ability to encode and retrieve spatial information.

Low receptive field sizes (e.g., r_{low} = 0.5, 1, 2) are suited for preserving spatial detail, which is particularly advantageous for far-apart stimuli. Our findings show that a higher inhibitory strength ($\beta = 0.3$) enhances this capability, leading to lower error rates, especially for smaller set sizes. The stronger inhibition appears to more effectively counteract excitatory spread, thereby minimizing overlap between neural representations and reducing encoding errors. This effect is particularly pronounced for far-apart stimuli under low receptive field conditions. In contrast, a lower inhibition level ($\beta = 0.2$) is associated with higher error rates when receptive fields are small. For close stimuli, the challenge of resolving tightly packed inputs persists with low receptive fields, as effective discrimination relies heavily on the precise balance of inhibition to prevent excessive pooling or interference.

Conversely, when employing high receptive field sizes (e.g., r_{high} = 4, 8, 16), the network exhibits increased capacity due to its ability to pool activity over larger areas. A key finding from our simulations is that, under these conditions, a *lower* inhibitory strength ($\beta = 0.2$) leads to improved performance with lower error rates across various stimulus distributions. This suggests that with broader receptive fields, stronger inhibition (such as $\beta = 0.3$) might unduly suppress widespread activity that could be necessary for representing objects or patterns spanning these larger fields. While high receptive fields inherently trade some spatial precision for capacity—leading to increased error rates for close stimuli due to greater overlap in neural activations—the reduced inhibition with $\beta = 0.2$ appears to strike a more effective balance for these larger fields. For far-apart stimuli, high receptive fields with this lower inhibition continue to perform reasonably well.

The distance-dependent inhibition, combined with noise, introduces variability that mimics biological systems. Figures 2 and 3 reveal that the optimal inhibitory strength β is not uniform but rather depends significantly on the receptive field size. As discussed, stronger inhibition ($\beta = 0.3$) is advantageous for low receptive field sizes, particularly for smaller set sizes and in resolving far-apart stimuli by minimizing representational overlap. In stark contrast, for high receptive field sizes, lower inhibition ($\beta = 0.2$) yields lower error rates. This highlights the critical importance of balanced excitation–inhibition dynamics, tailored to the network's structural properties like receptive field size, for achieving optimal performance. These findings align with the theoretical framework of spatially distributed memory systems.

The results of our simulation suggest that the capacity limitations observed in VWM tasks, particularly the rapid decline in performance beyond set sizes of 4-5 items, may not necessarily reflect a fixed number of memory “slots” as previously proposed by[23]. Instead, our findings provide an alternative explanation rooted in the computational constraints of a spatiotopic hierarchical processing network. Specifically, the performance limitations in working memory may arise due to encoding errors introduced by the overlap of receptive fields in higher layers of the network, where receptive fields become progressively larger.

In the change detection paradigm, subjects typically judge whether a test array differs from a memory array by a single item, with performance declining sharply beyond four to five items[24]. While this has traditionally been interpreted as evidence for discrete slots in working memory, our model suggests that this decline can be understood in terms of the network's spatial resolution constraints. As set sizes increase, and especially as items are presented closely together, the overlap of receptive fields (cross-talk) in higher layers of the network increases, leading to encoding errors that manifest as reduced accuracy (see Fig. 4).

Furthermore, the interaction between receptive field size, density, and spatial arrangement is critical for understanding how encoding errors emerge in the network. In regions of the visual field with high cortical magnification, such as the central portion of the visual field, higher receptive field density may compensate for the proximity of stimuli, allowing for improved capacity despite closely packed presentations. However, as stimuli are presented more eccentrically or in conditions where receptive field density decreases, the capacity to clearly represent multiple objects declines, and the network's performance is likely to degrade further due to increased cross-talk.

Our proposed hierarchical VWM model (see Algorithm 2) implements a multi-layer recurrent architecture that mirrors the processing hierarchy observed in the visual cortex. Table 1 summarizes the key simulation parameters, which are chosen to reflect biologically plausible conditions. Figure 6 illustrates the p-lattice organization—from the global scene representation to object-level encoding and the subsequent detailed feature layers—thereby establishing the topographic layout of the network. As shown in Fig. 7, recall probability decreases with increasing set size, indicating capacity limitations inherent in the system. Crucially, the cross-talk function $C(L)$ captures encoding errors arising from receptive field overlaps in the visual hierarchy. These errors dominate the network's performance when feedback is minimal; however, as recurrent localization becomes more extensive, the clarity of feature representations improves, reducing encoding errors and enhancing recall performance.

This model aligns with the computational theory proposed by[18], which highlights the inherent limitations of hierarchical networks in representing distinct spatial loci due to overlap in higher processing layers. Our simulation results indicate that the capacity bottleneck observed in VWM may stem from the network's inherent structural limitations, with receptive field sizes and densities constraining the number of distinct objects that can be clearly represented. Consequently, the degradation in change detection accuracy for set sizes larger than four may be driven by these encoding errors, rather than by a discrete number of memory slots.

In this work, we propose a hierarchical VWM model that mirrors the visual processing hierarchy (VH) in structure and function, with feedback loops that enable incremental refinement of visual information (See Fig. 5). Our model integrates abstract object-level representations in the Prefrontal Cortex (PFC) with lower-level feature representations that are processed through recurrent on-center off-surround (OCOS) dynamics. By modulating the degree of localization through feedback processing, the model allows only a partial, less-detailed representation at higher levels (IT or TEO layers) until recurrent processing reaches lower levels (V4 or V1), where fully localized features can be processed. This hierarchical, feedback-based approach reflects physiological findings that suggest a combination of bottom-up stimulus-driven processing and top-down attentional modulation in visual working memory[13,18] . The cross-talk function $C(L)$ in our model also plays a key role in gating the clarity and specificity of feature representations based on how deeply the feedback process has reached through VH, capturing a dynamic level of feature individuation.

Our hierarchical VWM model, which utilizes an on-center off-surround architecture with distance-dependent inhibition and top-down feedback via a cross-talk function, aligns with several recent findings in multi-item working memory research. Notably, studies employing rate attractor models in large networks[25,26] have demonstrated that sharpening excitatory or inhibitory interactions increases working memory capacity. Our simulation results similarly indicate that enhancing the precision of inhibition through distance-dependent mechanisms can reduce interference between closely spaced items, thereby increasing the effective capacity of the network. In our model, the cross-talk function modulates the clarity of feature representations as feedback deepens, effectively "sharpening" the responses at lower levels. This suggests that, like the attractor models, our framework captures an essential mechanism whereby refined inhibition contributes to improved memory fidelity and capacity. Future work will involve a quantitative comparison of our model's parameter space with those used in established attractor models to further clarify these similarities and to better situate our findings within the broader literature on multi-item working memory.

Compared to other VWM models, which often treat memory content as static or uniform across representations, our model emphasizes the graded refinement of visual information according to attentional engagement and feedback depth. For instance, models like the Slot Model[23] assume discrete storage limits without accounting for varying levels of feature clarity, while Resource-Based Models[27] distribute memory capacity as a continuous resource but do not fully capture the hierarchical interaction between different levels of abstraction and localization. Additionally, the Episodic Buffer Model[28] introduces integration across sensory inputs but lacks a structured hierarchy aligned with VH. More recently[29] have proposed high dimensional representations centered around PFC, where capacity limit arises from a balance of integration and interference. Our approach not only enhances the understanding of VWM's role in supporting attention-driven processing but also aims to clarify the content of VWM by aligning memory representations more directly with the hierarchical and recurrent processing found in the VH. This layered model highlights the importance of feedback and dynamic updating in VWM, suggesting that memory content is not static but is progressively refined in response to attentional demands and hierarchical feedback.

While our work emphasizes the role of recurrent on-center off-surround dynamics with distance-dependent inhibition in visual working memory (VWM), it is important to contextualize our findings within the broader landscape of modern computational modeling. Recent studies have demonstrated that sharpening excitatory and inhibitory interactions can robustly enhance the capacity and fidelity of VWM representations. For example,[30] showed using rate attractor models that NMDA-mediated recurrent excitation, when balanced with tuned lateral inhibition, stabilizes persistent activity and improves memory precision. Similarly, foundational models by[31,32] have illustrated how the interplay between excitation and inhibition can yield attractor-like memory states, with sharper inhibition leading to more precise representations. More recent computational frameworks, such as those proposed by[33] and[34], further demonstrate that mechanisms like fast Hebbian plasticity and neuromodulatory control are critical for maintaining stable, high-fidelity VWM representations.

In our model, the cross-talk function $C(L)$ is introduced to capture the encoding error due to overlapping receptive fields when only a fraction $A(L)$ of the network is effectively engaged. This function, expressed as $C(L) = A(L)\exp\left[-B\left(1 - A(L)\right)^2\right]$, provides a quantitative measure of the trade-off between network capacity and spatial precision. Unlike previous models that focus primarily on steady-state attractor dynamics, our approach integrates dynamic feedback processes, thereby addressing the variability in effective network coverage. By unifying these computational mechanisms into a single framework, our work offers novel insights into the conditions under which VWM capacity can be optimized, and it provides a basis for further investigations into the modulation of memory fidelity by recurrent feedback.

While our simulation results provide valuable insights into how recurrent on-center off-surround dynamics and distance-dependent inhibition can shape visual working memory (VWM), our model has several limitations that warrant further discussion. First, the current implementation employs a relatively small network (N=64, arranged in an 8×8 grid), which may not fully capture the complexity and variability of cortical processing in the brain. Furthermore, although our hierarchical model is loosely based on established visual processing hierarchies[13,15], the structure and connectivity patterns employed here are a significant simplification of the intricate, multi-area networks observed in vivo. In addition, the use of Gaussian receptive field profiles and uniform distance-dependent inhibition does not address known heterogeneity in cortical receptive field sizes and connectivity dynamics, as highlighted by more recent studies[30].

In future work, we aim to address these limitations by scaling our model to larger network sizes and incorporating more biologically realistic connectivity patterns that reflect the heterogeneity of receptive fields. Specifically, we plan to implement adaptive feedback mechanisms and neuromodulatory influences that can dynamically adjust the cross-talk between layers, thereby refining our model's predictive power in capturing interference effects and capacity limits.

Conclusion

In conclusion, we have presented a hierarchical model of visual working memory (VWM) that employs recurrent on-center off-surround dynamics with distance-dependent inhibition to simulate the encoding and maintenance of multiple objects. Our approach introduces a novel cross-talk function, which quantitatively captures the loss in effective stimulus clarity due to receptive field overlaps when only a fraction of the network is engaged. Simulation results demonstrate that while larger receptive field sizes result in lower overall error rates by enhancing global pattern matching, they may also lead to increased overlap and reduced spatial precision when stimuli are closely spaced. Conversely, smaller receptive fields offer higher resolution for spatially dispersed inputs. These findings underline a fundamental trade-off between capacity and precision in VWM. Our model, which builds on and extends previous attractor-based frameworks, offers new insights into how recurrent localization and feedback interactions can dynamically modulate memory fidelity. Future work will focus on scaling up the network to more biologically realistic connectomes and incorporating adaptive plasticity mechanisms to further refine these computational principles.

Data availability

The codes needed to run all the simulations can be found in https://github.com/rakesh-sengupta/vwm-change-detect/.

Received: 30 October 2024; Accepted: 11 June 2025
Published online: 02 July 2025

References

1. Usher, M. & Cohen, J. D. Heinke, D., Humphreys, G. W. & Olson, A. (eds) Short term memory and selection processes in a frontal-lobe model. (eds. Heinke, D., Humphreys, G. W. & Olson, A.). In *Connectionist Models in Cognitive Neuroscience*. 78–91 (Springer, 1999).
2. Vindhya, L. S., Gnana Prasanna, R., Sengupta, R. & Shukla, A. Nicosia, G. et al. (eds) Modeling primacy, recency, and cued recall in serial memory task using on-center off-surround recurrent neural network. (eds. Nicosia, G. et al.). In *Machine Learning, Optimization, and Data Science*. 405–414 (Springer, 2024).
3. Bhatnagar, R. *Edge Detection by Artificial Visual Cortex*. Ph.D. thesis, IIT Kharagpur (2012). https://cse.iitkgp.ac.in/~dsamanta/resources/thesis/Rajat-Bhatnagar-Thesis.pdf.
4. Verma, B. & Sengupta, R. Enumeration module for human visual system: A recurrent on-center off-surround neural network approach. *Academia* (2023). https://www.academia.edu/download/106187470/latest.pdf.
5. Ross, W. D., Grossberg, S. & Mingolla, E. Visual cortical mechanisms of perceptual grouping: Interacting layers, networks, columns, and maps. *Neural Netw.* **13**, 571–588 (2000).
6. Hansen, T., Sepp, W. & Neumann, H. *Recurrent Long-Range Interactions in Early Vision*. 127–138 (Springer, 2001). https://doi.org/10.1007/3-540-44597-8_9.
7. Thielscher, A. & Neumann, H. Neural mechanisms of cortico-cortical interaction in texture boundary detection. *Neuroscience* **122**, 647–660 (2003).
8. Marić, M. *Neurodynamic Models of Top-Down Effects on Visual Perception*. Ph.D. thesis, University of Rijeka (2022). https://dr.nsk.hr/islandora/object/ffri:3026.
9. Kurt, A. G. *Characterizing Surround Suppression in Motion Direction Perception*. Ph.D. thesis, Bilkent University (2021). https://repository.bilkent.edu.tr/bitstream/handle/11693/76419/10409440.pdf.
10. Öğmen, H. Spatiotemporal dynamics of visual perception across neural maps and pathways (ed.Corrochano, E. B.). In *Handbook of Geometric Computing: Applications in Pattern Recognition, Computer Vision, Neuralcomputing, and Robotics*. 1–29 (Springer, 2005). https://doi.org/10.1007/3-540-28247-5_1.
11. Sengupta, R., Surampudi, B. R. & Melcher, D. A visual sense of number emerges from the dynamics of a recurrent on-center off-surround neural network. *Brain Res.* **1582**, 114–124 (2014).
12. Thielscher, A. *Nonlinear Recurrent Mechanisms for the Processing of Surface Boundaries Based on Luminance and Texture Gradients*. Ph.D. thesis, University of Ulm (2008).

13. Bullier, J. Integrated model of visual processing. *Brain Res. Rev.* **36**, 96–107 (2001).
14. Lamme, V. A. & Roelfsema, P. R. The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* **23**, 571–579 (2000).
15. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202 (2001).
16. Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G. & Mishkin, M. The ventral visual pathway: An expanded neural framework for the processing of object quality. *Trends Cognit. Sci.* **17**, 26–49 (2013).
17. Sengupta, R., Bapiraju, S. & Pattanayak, A. Nicosia, G. et al. (eds) Exploring emergent properties of recurrent neural networks using a novel energy function formalism. (eds Nicosia, G. et al.). In *Machine Learning, Optimization, and Data Science*. 303–317 (Springer, 2024).
18. Tsotsos, J. K. *A Computational Perspective on Visual Attention* (MIT Press, 2011).
19. VanRullen, R. & Thorpe, S. J. The time course of visual processing: From early perception to decision-making. *J. Cognit. Neurosci.* **13**, 454–461 (2001).
20. Hegdé, J. Time course of visual perception: Coarse-to-fine processing and beyond. *Prog. Neurobiol.* **84**, 405–439 (2008).
21. Miller, E. K. The prefontral cortex and cognitive control. *Nat. Rev. Neurosci.* **1**, 59–65 (2000).
22. Kiyonaga, A. & Egner, T. Center-surround inhibition in working memory. *Curr. Biol.* **26**, 64–68 (2016).
23. Luck, S. J. & Vogel, E. K. The capacity of visual working memory for features and conjunctions. *Nature* **390**, 279 (1997).
24. Rouder, J. N., Morey, R. D., Morey, C. C. & Cowan, N. How to measure working memory capacity in the change detection paradigm. *Psychon. Bull. Rev.* **18**, 324–330 (2011).
25. Edin, F. et al. Mechanism for top-down control of working memory capacity. *Proc. Natl. Acad. Sci.* **106**, 6802–6807 (2009).
26. Wei, Z., Wang, X.-J. & Wang, D.-H. From distributed resources to limited slots in multiple-item working memory: A spiking network model with normalization. *J. Neurosci.* **32**, 11228–11240 (2012).
27. Bays, P. M. & Husain, M. Dynamic shifts of limited working memory resources in human vision. *Science* **321**, 851–854 (2008).
28. Baddeley, A. The episodic buffer: A new component of working memory?. *Trends Cognit. Sci.* **4**, 417–423 (2000).
29. Buschman, T. J. Balancing flexibility and interference in working memory. *Annu. Rev. Vis. Sci.* **7**, 367–388 (2021).
30. Murray, J. D., Jaramillo, J. & Wang, X.-J. Working memory and decision-making in a frontoparietal circuit model. *J. Neurosci.* **37**, 12167–12186 (2017).
31. Compte, A., Brunel, N., Goldman-Rakic, P. S. & Wang, X.-J. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cereb. Cortex* **10**, 910–923 (2000).
32. Wang, X.-J. Synaptic basis of cortical persistent activity: The importance of NMDA receptors to working memory. *J. Neurosci.* **19**, 9587–9603 (1999).
33. Fiebig, F. & Lansner, A. A spiking working memory model based on Hebbian short-term potentiation. *J. Neurosci.* **37**, 83–96 (2017).
34. Qi, X.-L. et al. Nucleus basalis stimulation enhances working memory by stabilizing stimulus representations in primate prefrontal cortical activity. *Cell Rep.* **36** (2021).

Author contributions

R.S. was responsible for the whole work including conceptualization, research, programming, analysis, visualization, drafting and editing of manuscript from initial to the final stages.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to R.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.