2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

ROBOVERINE: A human-inspired neural robotic process model of active visual search and scene grammar in naturalistic environments

Raul Grieben[1], Stephan Sehring[1], Jan Tekülve[1], John P. Spencer[2] and Gregor Schöner[1]

***Abstract*— We present ROBOVERINE, a neural dynamic robotic active vision process model of selective visual attention and scene grammar in naturalistic environments. The model addresses significant challenges for cognitive robotic models of visual attention: combined bottom-up salience and top-down feature guidance, combined overt and covert attention, coordinate transformations, two forms of inhibition of return, finding objects outside of the camera frame, integrated space- and object-based analysis, minimally supervised few-shot continuous online learning for recognition and guidance templates, and autonomous switching between exploration and visual search. Furthermore, it incorporates a neural process account of scene grammar — prior knowledge about the relation between objects in the scene — to reduce the search space and increase search efficiency. The model also showcases the strength of bridging two frameworks: Deep Neural Networks for feature extractions and Dynamic Field Theory for cognitive operations.**

I. INTRODUCTION

Most goal-oriented interactions with the environment entail a preceding visual search. Effective feature guidance [1] helps reduce the number of saccades needed to find the target object in a scene, and the combination of overt and covert attention shifts [2] allows us to scan complex scenes efficiently despite the visual system's limitations. Natural scenes tend to be cluttered but highly structured, and humans use their knowledge about the relation between objects in scenes - the scene grammar [3] - to reduce the search space. Importantly, humans are not limited to finding objects they already know. Cognitive robotics aims to develop autonomous agents with cognitive abilities similar to humans (see [4] for a recent overview of the state-of-the-art in human-inspired robotic vision). Begum and Karray [5] created a list of issues and challenges that a cognitive model for robot attention needs to address:

1) Combine covert and overt modes of attention.
   1.1) Cope with at least four coordinate systems: world, head, camera, and image coordinates.
   1.2) Integrate space-based and object-based inhibition of return.
   1.3) Cope with objects leaving the camera frame or being only partially visible due to head movement.
2) Integrate space- and object-based analysis.
3) Learn an object's visual features with minimal human supervision from different view angles.
4) Autonomous switching between exploration and visual search based on the task.
5) Learning while working, without needing a separate training phase (online learning).

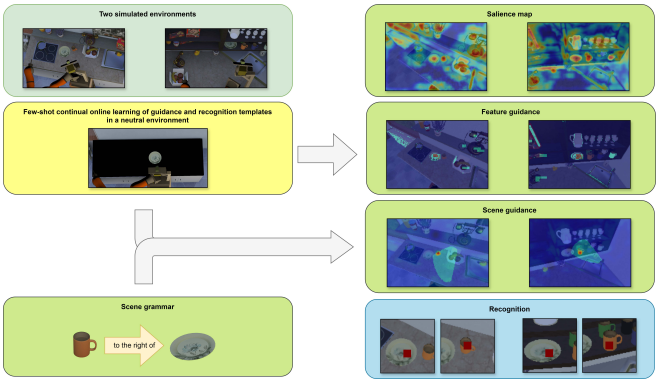

Fig. 1. A simplified overview of the problem (left) and the cognitive operations needed to solve it (right).

Here, we present ROBOVERINE, a neural robotic process model that addresses these issues (Figure 1) building upon our previous work on human attention ([6], [7]). We show that it can control an autonomous agent in different simulated environments. Furthermore, we also included a neural process account of scene grammar (see [8] for a related approach). Interfacing the neural architecture based on Dynamic Field Theory (DFT; [9]) with a pre-trained headless convolutional neural network (CNN; VGG16: [10]) for feature extraction was necessary to enable interaction with natural scenes. The interface is based on neurally plausible learning and combines the two frameworks' strengths. DFT delivers autonomous process organization, sequence generation, and working memory. The CNN extracts the complex features needed for object recognition. We use the Bienenstock-Cooper-Munro (BCM) rule [11] to learn the mapping from the distributed representation of the CNN feature maps to the localist representation of a 3D neural field defined over space and visual category. This localist representation over parafoveal space enables the cognitive operation of attentional selection around the current fixation point, combining overt and covert attention in the model. Importantly, this also allows for continually learning object recognition templates for new classes and guidance templates that support visual search for the new classes. Such online learning is essential for efficient human-robot interaction [5].

This work was supported by the Leverhulme Trust Research Project Grant (RPG-2021-350). (Raul Grieben and Stephan Sehring are co-first authors.)

[1]Faculty of Computer Science, Institute for Neural Computation, Ruhr University Bochum, Bochum, Germany {raul.grieben, stephan.sehring, jan.tekuelve, gregor.schoener}@ini.rub.de

[2]School of Psychology, University of East Anglia, Norwich, United Kingdom j.spencer@uea.ac.uk

II. METHODS

The neural process model is based on Dynamic Field Theory (DFT; [9]), a mathematical framework that aims to understand how cognition emerges from neural population activation and discrete events that arise from instabilities in the underlying dynamics. The time-continuous evolution of graded activation patterns, $u(x,t)$, of a neural population, tuned to a dimension x, is formalized as a *dynamic neural field*. The activation $u(x,t)$ changes over time t, on the time scale τ, according to the integro-differential equation [12]:

$$\tau\dot{u}(x,t) = -u(x,t) + h + s(x,t) + w\xi(x,t) + \int \omega(x-x')\ \sigma(u(x',t))\ dx' \quad (1)$$

As long as the activation is below the threshold of the sigmoidal function $\sigma(u) = 1/(1+\exp[-\beta u])$ the system's attractor state is established by the stabilizing term, $-u(x,t)$, the negative resting level, h, and the external input, $s(x,t)$. Above threshold activation engages lateral interactions defined by the kernel, $\omega(x-x')$, that combines local excitation and global inhibition leading to the formation of self-stabilized supra-threshold activation peaks. Such peaks, the units of representation in DFT, are induced when increasing input pushes sub-threshold activation states through the *detection instability*. Peaks disappear when decreasing input drives them through the *reverse detection instability*. The two instabilities delimit a bistable regime. Weighted Gaussian white noise, $w\xi(x,t)$, induces fluctuations that enable switching between stable states near instabilities.

Different dynamic regimes arise for varying kernel parameters. In the *self-stabilized* regime, peaks resist fluctuations in input strength due to bistability. In the *selective* regime, a peak can only be formed over one location, bringing about a selection decision when multiple sites received localized input. In the *sustained activation* regime, peaks persist after removing localized input completely, a neural implementation of working memory.

Networks of fields are build by coupling different fields, including dimension expanding and dimension contracting coupling patterns. Cognitive and motoric processes emerge from such networks. Together with the dynamic instabilities, such networks enable sequences of discrete processing steps.

III. MODEL

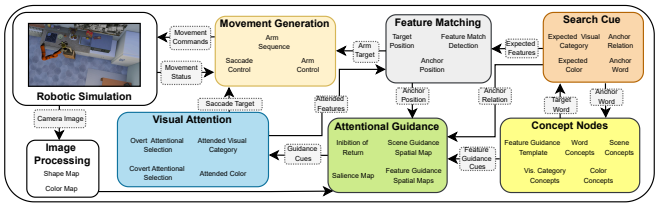

Fig. 2. An overview of the neural dynamic process model.

The neural dynamic process model, depicted in a simplified outline in Figure 2, autonomously controls a robot to visually explore its environment driven by salience, and, in the presence of a search cue, to perform active visual search for real-world object categories in naturalistic scenes. It keeps the location of a found target in working memory, enabling object-oriented behavior such as grasping. Search efficiency is increased by using the known semantic structure of the scene, the scene grammar, to bias search toward locations that are in an appropriate spatial relation to a detected anchor object [3]. The model will also find objects whose location violates the expectation of the scene grammar, and will learn new guidance and recognition templates for novel object categories in continual few-shot online learning.

The full model shown in Figure 3 is a set of coupled integro-differential equations, in which neural activation evolves continuously in time. Events and transitions arise at discrete moments in time from instabilities in the dynamics. Various cognitive operations emerge from this neural dynamics. The labels in the figure and the terms we use to describe the operation of the model below refer to the functional significance of time courses of activation in different subnetworks (referenced by uppercase letters) and component neural fields. The model was implemented and numerically simulated in *cedar* [13]. The robotic agent and the simulated environments were created and simulated using *Webots* [14].

A. Feed-forward feature and salience maps

The model's bottom-up pathway simultaneously extracts low-resolution retinal (D) and higher-resolution parafoveal (I) features from the camera image (A) through a parallel preattentive process.

1) Functional visual fields: Taking inspiration from the human attentional system, we incorporated neural processes for three types of *functional visual fields* (FVF) [2]: the *exploratory FVF*, the *attentional FVF*, and the *resolution FVF*. The *exploratory FVF* is responsible for overt attention shifts by selecting the next fixation location in it's retinotopic subspace guided by low-resolution coarse feature information. In contrast, the *attentional FVF* is the central parafoveal area around the fixation point within which covert attention shifts are possible. The *resolution FVF* is the area within which highly detailed feature information can be extracted, allowing objects to be identified covertly. For simplicity, we assume that *resolution* and *attention FVFs* are the same in the model.

2) Exploratory FVF feature extraction: Preattentive *color* and preattentive *shape* are extracted from a scaled-down version (120 x 180 pixel) of the camera image (299 x 448 pixel). Preattentive color is extracted from hue-space ($D3$). Preattentive shape is extracted from the intermediate *conv 4-3* convolutional layer of the VGG16 network ($D4$). Each *space/feature map* field (E) receives input from the corresponding feature stack ($D1, D2$), and their activation is marginalized along the feature dimension by using a center-surround filter ($H1$) as the projection kernel, resulting in one feature *conspicuity map* ($K3$) for each feature. These serve as input to the *salience map* ($K1$) [15] of the model.

3) Attentional FVF feature extraction: From the original camera image (299 x 448 pixels), *color* and *visual category*

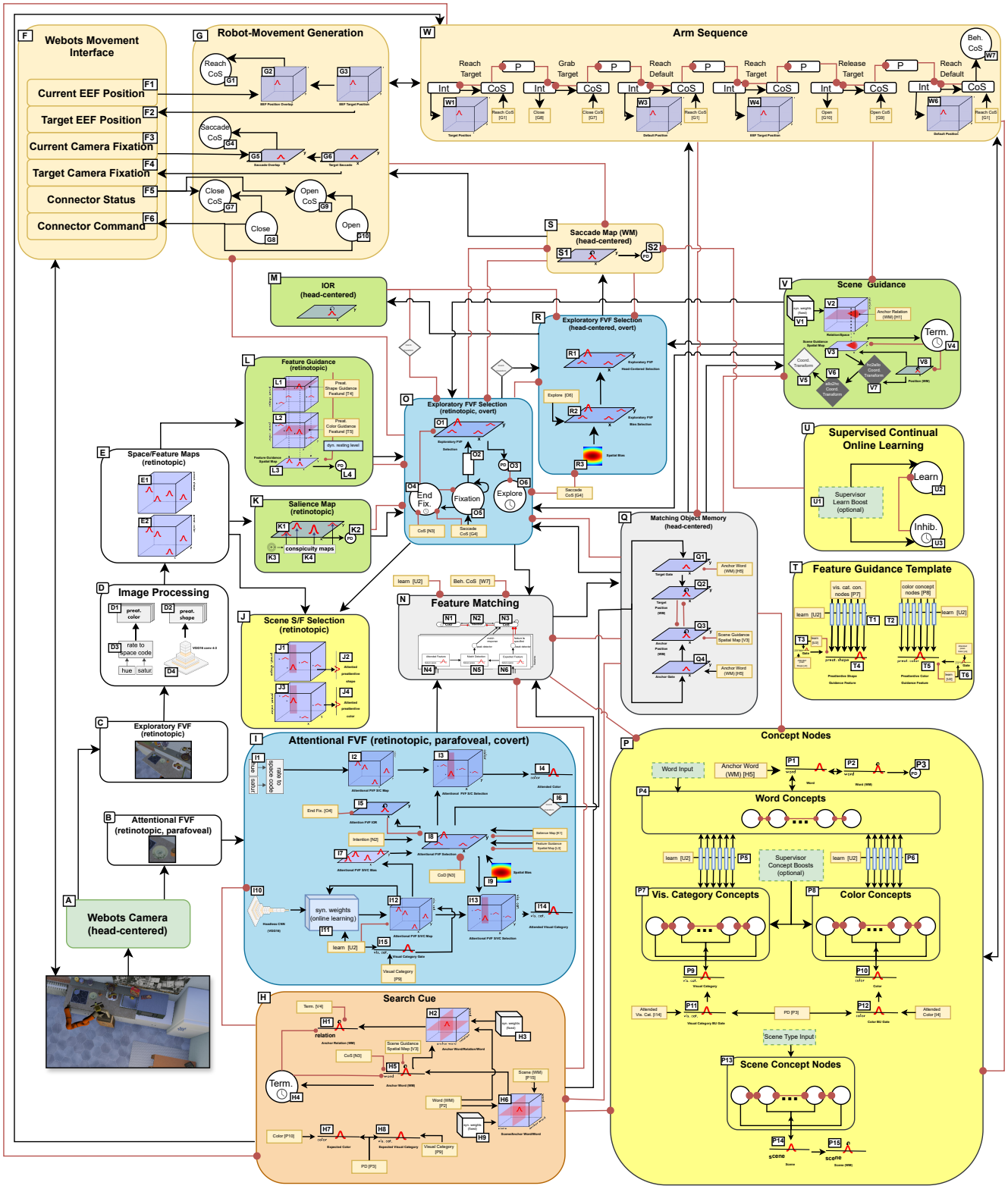

Fig. 3. The neural dynamic process model.

are extracted from the *attentional FVF* image (*B*), taken from the center (80 x 80 pixels). Color is extracted from hue-space (*I*1) and serves as input to the *attentional FVF space/color map* (*I*2). Visual category is the result of a learned mapping from the last convolutional layer (*conv 5-3*) of the headless VGG16 CNN (*I*10) to the *attentional FVF space/visual category map* (*I*12).

B. Attentional selection

Visual selective attention is foundational to flexible, goal-oriented human behavior. In the model, all visual cognitive processes result from attentional selection.

1) Exploratory FVF attentional selection (overt attention): The *exploratory FVF selection* field (*O*1) selects a location to fixate next in the retinotopic space of the *exploratory FVF* through biased competition [16]. It receives an excitatory bottom-up bias from the *salience map* (*K*1), and two excitatory top-down guidance biases from the *feature guidance spatial map* (*L*3) and the coordinate transformed *scene guidance spatial map* (*V*3) fields. Through its connection pattern, the *fixation* node (*O*5) induces a peak in the center of the *exploratory FVF selection* field (*O*1) when the *saccade CoS* node (*G*4) signals the completion of a saccade. This peak prevents the selection of another location until the peak becomes de-activated by the *end fixation* node (*O*4) that operates on a slower timescale. The *inhibition of return* field (IOR; *M*) biases attention away from previously attended locations, enabling sequences of selection decisions. If no selection decision is made in the *exploratory FVF selection* field (*O*1), during a fixed time window, the *explore* node (*O*6) will become active and drive an exploratory saccade in the head-centered space by enabling peak formation in the *exploratory FVF bias selection* field (*R*2). The *exploratory FVF selection* field (*O*1) serves as coordinate transformed input to the *exploratory FVF head-centered selection* field (*R*1) that drives the saccadic system by inducing a peak in the *saccade map (WM)* (*S*1)),

2) Attentional FVF attentional selection (covert attention): The *attentional FVF selection* field (*I*8) selects a location in the *attentional FVF* to be attended covertly. The *intention* node (*N*2) and the *condition of dissatisfaction* node (CoD; *N*1) constitute a neural oscillator [12]. The *intention* node (*N*5) homogeneously boosts the *attentional FVF selection* field (*I*8), allowing the field to make a selection decision. That field is homogeneously inhibited by the *CoD* node (*N*1), destabilizing any peak that has built. This field receives excitatory bottom-up bias input from the *salience map* (*K*1) and excitatory top-down guidance bias from the *feature guidance spatial map* (*L*3). Further, it receives a spatial bias that favors the center of the *attentional FVF* and the *attentional FVF space/visual category bias* (*I*7) that increases the probability of selecting a location that is associated with a known visual category. This bias results from marginalizing the activation of the *attentional FVF space/visual category map* (*I*12) along the visual category dimension. The *attention FVF IOR* (*I*5) guides attention away from covertly attended locations and is homogenously inhibited at the end of a fixation. At each covertly attended location, the *attended color* (*I*4) and *visual category* (*I*14) are extracted through selection in the corresponding *attentional FVF space/feature selection* field where input from the corresponding *attentional FVF space/feature map* overlaps with the localized input from the *attentional FVF selection* field (*I*8).

C. Visual search

Visual search requires the neural activation in working memory of a guidance template for the target object [1] that biases selection through top-down feedback loops. A peak in the *target position (WM)* field (*Q*2) terminates the visual search.

1) Search cue: A simulated language interaction (e.g. "Look for a cup in the kitchen") triggers the model's visual search by activating a word (*P*4) and a scene concept node (*P*13). The selective *scene/anchor word/word)* field (*H*6) is a simplified long-time memory (LTM) representation of learned *scene/anchor word/word* combinations. The same is true for the *anchor word/relation/word)* field (*H*2). If an association between the current scene and word exists, a peak is formed in the *anchor word (WM)* field (*H*5), which then becomes the new target of the current visual search by providing input to the *word* field (*P*1) which activates the corresponding word concept node through bidirectional coupling. The model autonomously switches back to the original target if no anchor object is found in the scene. If an association between the current anchor word and word exists, a peak is formed in the *anchor relation (WM)* field (*H*1). An active word concept node (*P*4) activates the associated visual category (*P*7) and color concept nodes (*P*8) through bidirectional Hebbian connections. This ultimately leads to corresponding peaks in the *expected visual category* (*H*8) and *color* (*H*7) fields that represent the current search cue.

2) Feature matching: Matches between *expected* (*H*7, *H*8) and *attended* (*I*4, *I*14) features are detected through the corresponding *match detection* fields (*N*5). If all features match the expected values, the *condition of satisfaction* node (CoS; *R*5) [17] is activated. Depending on the state of the *anchor word (WM)* (*H*5), the currently attended location is committed to the *anchor position* (*Q*3) or the *target position* (*Q*2) working memory.

3) Feature guidance: An active visual category concept node (*P*7) induces a peak in the *preattentive shape guidance feature* field (*T*4) through Hebbian-learned connection weights. The same is true for a color concept node (*P*8) and the *preattentive color guidance feature* field (*T*5) These fields give input to the *feature guidance cue* fields (*L*1 and *L*2). Peaks from at locations where this input overlaps with the input from the corresponding *space/feature map* field (*E*1 and *E*3). Activation in these fields is marginalized along the feature dimension and serves as input to the *feature guidance spatial map* (*L*3), whose resting level is dynamically down-regulated through inhibitory coupling.

4) Scene guidance/grammar: The model reduces the search space by using anchor objects and their known

spatial relation to other objects in the scene. To provide attentional bias relative to a found anchor object, the model uses a set of coordinate transformations [18] of activation patterns that represent operators in relational spatial language [19]. This spatial pattern is formed in the *relation/space* field ($V2$) through an overlap between the *anchor relation (WM)* ($H1$) and the synaptic relation connection patterns. The marginalized activation along the relation dimension is input to the *scene guidance spatial map* ($V3$), and this field is inhibited if an anchor object is found or after enough time has elapsed through homogeneous inhibition from the *termination* node ($V4$). The position of the anchor object ($V6$) is first coordinate transformed from head-centered to allocentric table space and then used to transform the scene guidance peak into allocentric space. The allocentric scene guidance peak is then transformed back to head-centered and retinotopic space to serve as input bias to the corresponding exploratory FVF selection fields ($O1$ and $R1$).

D. Learning

Autonomous learning was implemented respecting neural plausibility. All non-fixed synaptic weights in the model are adapted by different variants of the Hebbian learning rule during learning. Since the model learns online and continuously, learning periods must be restricted to meaningful events. Learning is initiated, therefore, by a learn boost ($U1$) that triggers an active transient activation pattern ($U2$,$U3$) [20]. A single such transient is sufficient to learn a new object class. The object is moved around in the attentional FVF during learning. During the transient learning phase, all plastic connections in the model adapt in parallel according to their respective update rules.

1) Learning of an object classification template: Complex feature maps m_f (where f is the feature index) are extracted from the *attentional FVF* image (B) through a pre-trained headless VGG16 network ($I10$). The object classification template consists of connection weights, $w_{m_\mathrm{f},u_\mathrm{afsv}}$, that perform the transformation from the distributed representation in the feature maps, m_f, to the localist representation in the *attentional FVF space/visual category map*, ($I12$) (u_afsv, $J5$). This mapping preserves spatial information to some degree, enabling the model to classify multiple objects in the *attentional FVF* through covert attention shifts similar as in human attention.

These connections weights are updated according to a dynamic version of the BCM ([11]) rule:

$$\begin{gathered}\tau_w \dot{w}_{m_\mathrm{f},u_\mathrm{afsv}}(x,t) = \eta \cdot \sigma(u_\mathrm{learn}(t)) \cdot y \cdot (y-\Theta) \cdot \frac{m_\mathrm{f}(x_1,x_2,t)}{\Theta} \\ y = \sigma(u_\mathrm{afsv}(x,t)) \\ \tau_\Theta \dot{\Theta} = (y^2 - \Theta),\end{gathered} \tag{2}$$

where η is the learning rate and u_learn is the transient activation of the *learn* node ($U2$). Before learning starts, the *visual category* field ($P9$) provides input to the *attentional FVF space/visual category map* ($I12$) through the *visual category gate* field ($I15$). This enables the association between a visual category concept node ($P7$) and the object classification template. At the same time, the *learn* node ($U2$) down-regulates through homogeneous inhibition the resting level of the *attentional FVF space/visual category map* ($I12$). After learning, the input s_{u_afsv} to the *attentional FVF space/visual category map* ($I12$) field is:

$$s_{u_\mathrm{afsv}}(x,t) = \sum_{f=0}^{F-1} m_\mathrm{f}(x_1,x_2,t) \cdot w_{m_\mathrm{f},u_\mathrm{afsv}}(x,t). \tag{3}$$

2) Learning of the guidance templates: The synaptic weight pattern, $w_{\mathrm{psgf_{vc}}}$, between visual category concept nodes (u_vc, $P7$) and the *preattentive shape guidance feature* field (u_psgf, $T4$) is updated according to a dynamic version of the Hebbian learning rule [21]:

$$\begin{gathered}\tau \dot{w}_{\mathrm{psgf_{vc}}}(x,t) = \eta \cdot \sigma(u_\mathrm{learn}(t)) \cdot \sigma(u_\mathrm{vc}(t)) \\ \cdot (\sigma(u_\mathrm{psgf}(x,t)) - w_{\mathrm{psgf_{vc}}}(x,t))\end{gathered} \tag{4}$$

The synaptic weight pattern, $w_{\mathrm{pcgf_c}}$, between color concept nodes (u_c, $P7$) and the *preattentive color guidance feature* field (u_pcgf, $T5$) is updated according to an analogous rule:

$$\begin{gathered}\tau \dot{w}_{\mathrm{pcgf_c}}(x,t) = \eta \cdot \sigma(u_\mathrm{learn}(t)) \cdot \sigma(u_\mathrm{c}(t)) \\ \cdot (\sigma(u_\mathrm{pcgf}(x,t)) - w_{\mathrm{pcgf_c}}(x,t))\end{gathered} \tag{5}$$

To enable association, the currently attended retinotopic features ($J2$ and $J4$) are gated during learning as inputs to the corresponding guidance feature field ($T4$ and $T5$).

3) Learning of word concepts: Word concepts are learned through co-activation. For this purpose, word input has to be provided during learning. The connection strength, $w_{\mathrm{w_{vc}}}$, between a word concept node (u_w, $P4$) and a visual category concept node (u_vc, $P7$) is updated according the dynamic Hebbioan learning rule:

$$\begin{gathered}\tau \dot{w}_{\mathrm{w_{vc}}}(t) = \eta \cdot \sigma(u_\mathrm{learn}(t)) \cdot \sigma(u_\mathrm{w}(t)) \\ \cdot (\sigma(u_\mathrm{vc}(t)) - w_{\mathrm{w_{vc}}}(t))\end{gathered} \tag{6}$$

The connection strength, $w_{\mathrm{w_c}}$, between a word concept node (u_w, $P4$) and a color concept node (u_c, $P8$) is updated according an analogous rule:

$$\begin{gathered}\tau \dot{w}_{\mathrm{w_c}}(t) = \eta \cdot \sigma(u_\mathrm{learn}(t)) \cdot \sigma(u_\mathrm{w}(t)) \\ \cdot (\sigma(u_\mathrm{c}(t)) - w_{\mathrm{w_c}}(t))\end{gathered} \tag{7}$$

E. Saccade and movement generation

The formation of a self-sustained peak in the *Saccade Map (WM)* field ($S1$) initiates a saccade towards the selected location. This peak inhibits the *Exploratory FVF Selection* and *Exploratory FVF Head-Centered Selection* fields ($O1$, $R1$), thus preventing new selection decisions while a saccade is ongoing. The *Target Saccade* field ($G6$) transmits the selected position to the simulated robot (*Target Camera Fixation* ($F4$)). The *Saccade CoS* node ($G4$) signals a successful saccade once the *Saccade Overlap* field ($G5$) detects the overlap of the *Current Camera Fixation* ($F3$) and the *Target Saccade* field. This inhibits the *Saccade Map (WM)* field and activates the *Fixation* node ($O5$), thereby generating a fixation peak.

The *Arm Sequence* (*W*) sub-network directs the robot arm to grab the target object. It can be seen as a placeholder for more sophisticated behavioral control structures [22] and is intended to show that information about a searched object can be used for further object-oriented behavior. The sub-network implements a sequence of *reach* and *open/close* movements. Each movement primitive is controlled by a pair of *Intention* (*I*) and *Condition of Satisfaction* (*CoS*) nodes that are organized into a sequence using *Precondition* (*P*) nodes [23]. Each *Intention* passes a movement command to the robot controller via the *EEF Target Position* field (*G*3) and *open/close* nodes (*G*8, *G*10). The *Reach CoS* and *open/close CoS* nodes (*G*1, *G*7, *G*9) signal the completion of a movement and activate the corresponding *CoS* node (*W*) of the behavioral sequence. Once the movement sequence is complete, the *Behavior CoS* node (*W*7) resets the model's memory, at which point a new word can be given as input.

IV. RESULTS

To showcase the agent's capabilities, we taught it to identify four object categories: plate, cup, telephone, and banana. Subsequently, we created two environments substantially distinct from each other and from the environment in which the agent learned the objects. Additionally, we constructed a basic scene grammar with two rules: the cup is to the right of the plate, and the banana is below the telephone. In the demonstrations shown here, the agent operates in the identity mode, in which it looks for the object that matches the one presented during the learning phase (for example, after learning "cup", it will look for an orange-colored object with a cup shape). The plots show snap shots of activation in selected dynamic neural fields. Complete activation time courses for these fields are available in the supplementary video.

A. Environment 1: Finding the cup and the banana

In the first task (Figure 4), we demonstrate that the agent can use feature and scene guidance to find learned objects in a complex simulated environment. At time 0.04s, the agent receives the command to find the cup (blue dotted line in the nodes plot). The cup word node becomes active first, and shortly after, a switch occurs, and the plate word node (the anchor object for the cup) becomes active. At time 0.5s, the feature guidance highlights the plate in the scene, and this location is selected in the exploratory FVF as a target for the next saccade. At 0.58s, the plate is covertly attended and recognized in the attentional FVF, which causes a switch back to the cup word node being active, the activation of the cup feature guidance, and a relational bias centered on the found plate location as input to the exploratory FVF. At 0.88s, the location of the cup is selected as the location for the next saccade because its feature match and the scene guidance having the highest activation. At 0.96s, the cup was covertly attended to and recognized after a saccade to its location, and at 1.25s, the agent grasped the found cup. The second half of the plot shows the same for the banana. It is important to note that this demonstrates how the agent can process new commands during live operation.

B. Environment 2: Finding the cup and the banana

The second demonstration (Figure 5) demonstrates that the agent can perform the same task as in the first demonstration despite the environment being significantly different. Since the time course are similar to those in the first demonstration, we only highlight noteworthy differences. At time 1.25s, we see the telephone, the banana's anchor object, outside the exploratory FVF image. As a result, the agent autonomously performs salience-driven exploratory saccades until the telephone becomes visible again. Through feature guidance, its location is selected as the target of the next saccade at time 2.7s. At 3.1s, the banana feature guidance failed to highlight the existing banana, so that there is no overlap between the banana and the relational bias. As a result, the next saccade does not fixate on the banana directly but on a location near it. The agent is still able to attend and recognize the banana covertly in the attentional FVF at time 3.32s. Without the relation guidance from the scene grammar, the agent would have needed more saccades to find the banana.

C. Environment 1: Finding the misplaced banana

In the third demonstration (Figure 6), we show that the agent can find the banana although it is not at the expected location. Specifically, at time 1.33s, the relational bias has vanished. The following saccade location is selected through a combination of feature guidance, salience, and IOR that favors the location of the banana.

V. DISCUSSION

We presented ROBOVERINE, a neural dynamic robotic process model that performs active guided visual search in naturalistic environments. It combines bottom-up salience and top-down feature guidance and incorporates overt and covert attention, coordinate transformations, and two types of inhibition of return. It performs integrated space- and object-based analysis and can learn new object classes with minimal supervision. Additionally, it autonomously switches between exploration and visual search and incorporates a neural process account of scene grammar. The model combines DNNs for feature extractions and DFT for cognitive operations. DNNs extract relevant features from the visual field, while DFT provides a robust framework for cognitive operations like attentional selection, autonomous learning, decision-making, autonomous process organization, sequence generation, and working memory. Our model has significant advantages over an end-to-end learned DNN in that it operates in a closed behavioral loop. The model's stable memory representations enable goal-oriented actions, while its adaptive recurrent top-down feedback allows top-down inference processes to switch between modes flexibly without requiring specific algorithms. Cognitive operations, like selection, require localist representations along the feature dimension. Attentional selection, therefore, requires a localist anatomically bound representation of features over

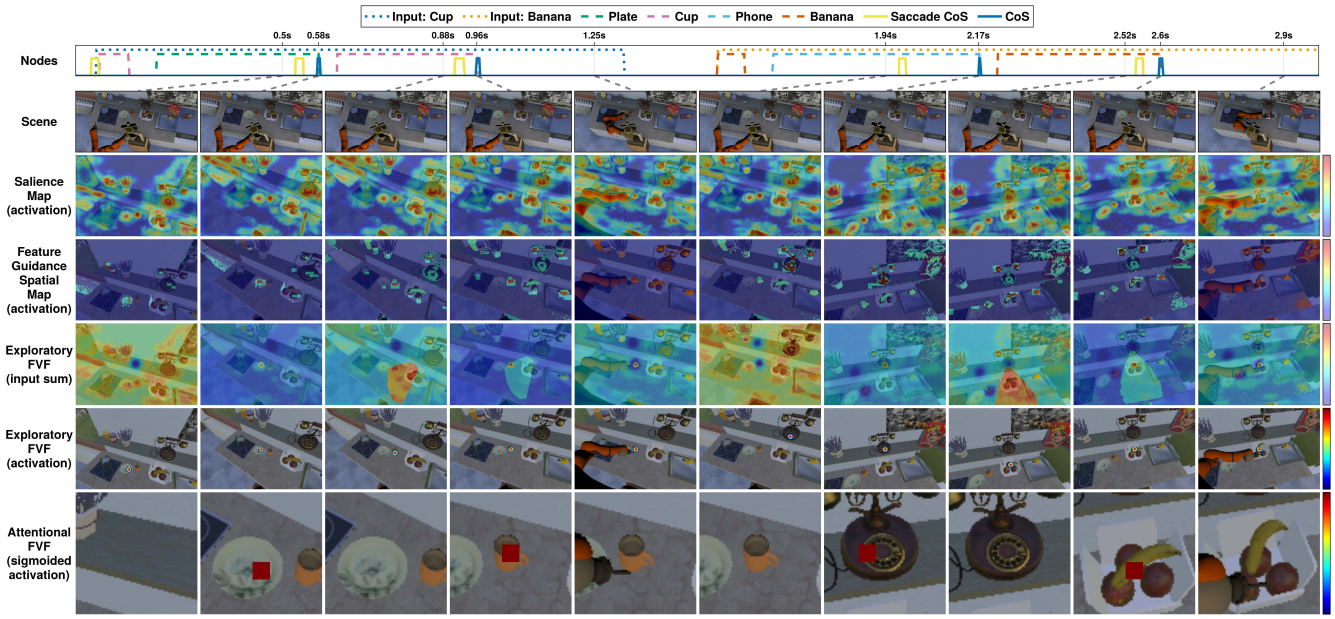

Fig. 4. Demonstration of the agent looking for the cup and the banana in the first kitchen (see text for an explanation).

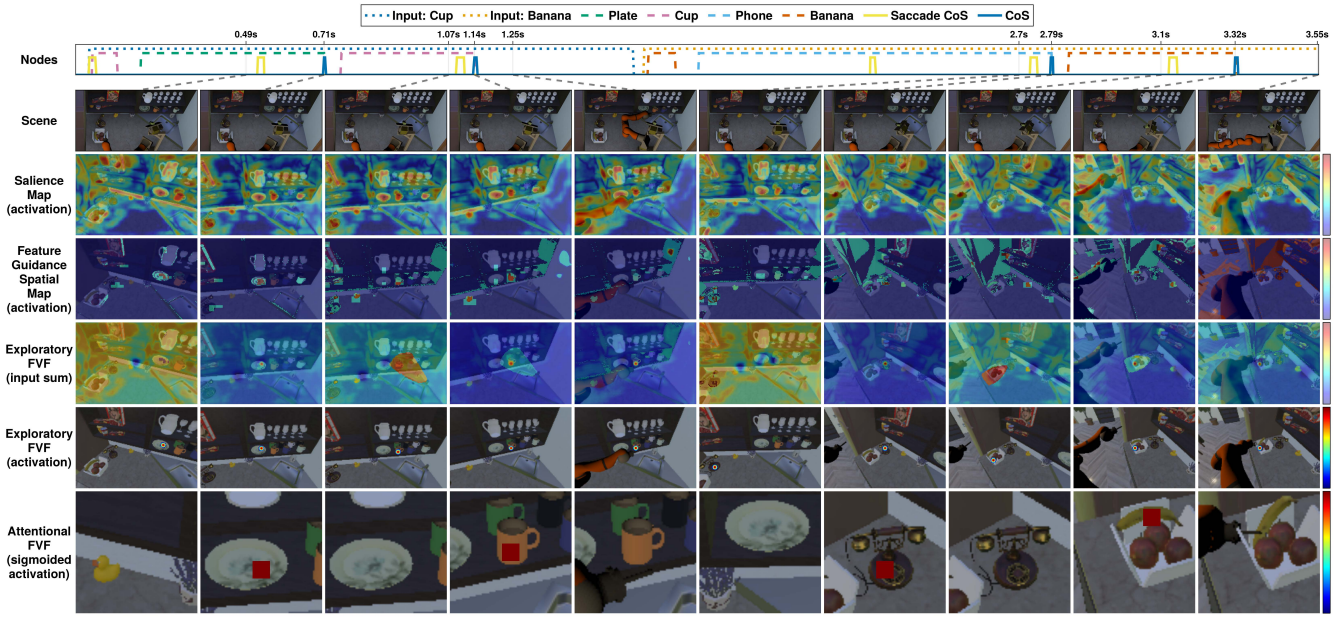

Fig. 5. Demonstration of the agent looking for the cup and the banana in the second kitchen (see text for an explanation)

space. Neural populations in the inferior temporal cortex (IT) represent object identity over space [24]. We suggested how to learn a mapping from the distributed feature representations of a CNN to a localist 3D neural field that enabled covert attentional selection and object recognition in the attentional FVF. For demonstration purposes, we only learned a small number of object categories, but the model does not have constraints that would prevent it from scaling to a large number of categories. The model aims to provide a pervasively neural process account of robotic active vision inspired by human visual attention. Therefore, its performance is hard to access using classical benchmarks since they do not cover its main innovations. While visual attention has been actively studied for many decades in the robotics context, a strong focus was on the bottom-up path of attention [25], [26]. We addressed some of the major active challenges in this research area [26], [5] by integrating bottom-up salience, task-driven top-down attention, and covert and overt attention in one time-continuous model. Future research involves optimizing the model's execution speed for real-time performance. This could be achieved by using optimized hardware or by replacing parts of the model

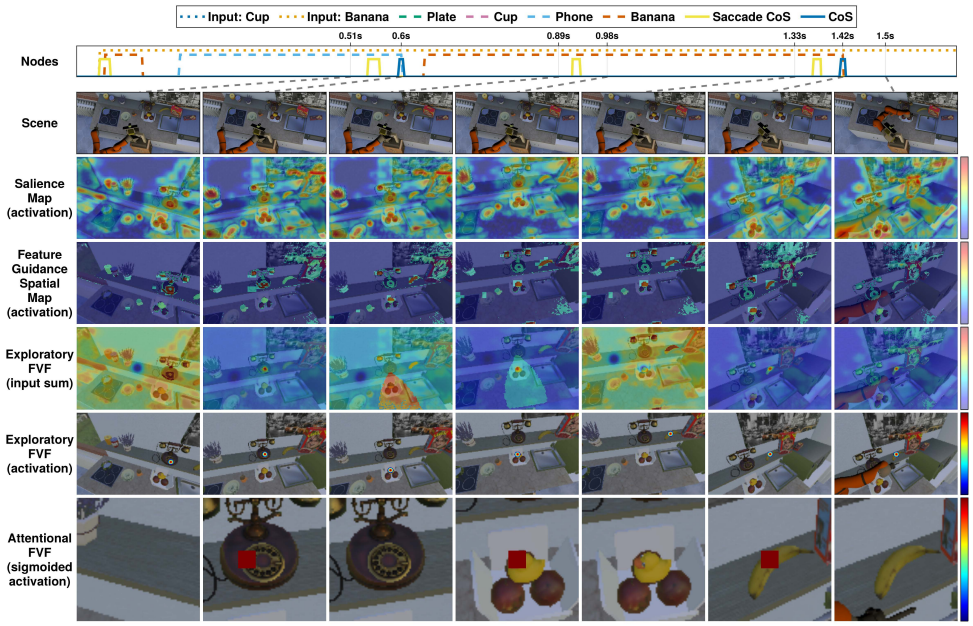

Fig. 6. Demonstration of the agent finding the misplaced banana in the first kitchen (see text for an explanation).

with analog algorithmic shortcuts.

REFERENCES

[1] J. M. Wolfe, "Guided search 6.0: An updated model of visual search," *Psychonomic Bulletin & Review*, pp. 1–33, 2021.

[2] C.-C. Wu and J. M. Wolfe, "The functional visual field (s) in simple visual search," *Vision Research*, vol. 190, p. 107965, 2022.

[3] M. L.-H. Võ, "The meaning and structure of scenes," *Vision Research*, vol. 181, pp. 10–20, 2021.

[4] R. P. de Figueiredo and A. Bernardino, "An overview of space-variant and active vision mechanisms for resource-constrained human inspired robotic vision," *Autonomous Robots*, vol. 47, no. 8, pp. 1119–1135, 2023.

[5] M. Begum and F. Karray, "Visual attention for robotic cognition: A survey," *IEEE Transactions on Autonomous Mental Development*, vol. 3, no. 1, pp. 92–105, 2010.

[6] R. Grieben, J. Tekülve, S. K. Zibner, J. Lins, S. Schneegans, and G. Schöner, "Scene memory and spatial inhibition in visual search: A neural dynamic process model and new experimental evidence," *Attention, Perception, & Psychophysics*, 2020.

[7] R. Grieben and G. Schöner, "Bridging DFT and DNNs: A neural dynamic process model of scene representation, guided visual search and scene grammar in natural scenes," in *Proceedings of the 44th Annual Conference of the Cognitive Science Society*, J. Culbertson, A. Perfors, H. Rabagliati, and V. Ramenzoni, Eds., 2022.

[8] A. Aydemir, K. Sjöö, J. Folkesson, A. Pronobis, and P. Jensfelt, "Search in the real world: Active visual object search based on spatial relations," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 2818–2824.

[9] G. Schöner, J. P. Spencer, and T. DFT Research Group, *Dynamic thinking: A primer on dynamic field theory*. Oxford University Press, 2016.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] C. C. Law and L. N. Cooper, "Formation of receptive fields in realistic visual environments according to the bienenstock, cooper, and munro (bcm) theory," *Proceedings of the National Academy of Sciences*, vol. 91, no. 16, pp. 7797–7801, 1994.

[12] S.-i. Amari, "Dynamics of pattern formation in lateral-inhibition type neural fields," *Biological cybernetics*, vol. 27, no. 2, pp. 77–87, 1977.

[13] O. Lomp, M. Richter, S. K. U. Zibner, and G. Schöner, "Developing Dynamic Field Theory Architectures for Embodied Cognitive Systems with cedar," *Frontiers in Neurorobotics*, vol. 10, p. 14, 2016.

[14] O. Michel, "Webots: Professional mobile robot simulation," *Journal of Advanced Robotics Systems*, vol. 1, no. 1, pp. 39–42, 2004.

[15] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision research*, vol. 40, no. 10-12, pp. 1489–1506, 2000.

[16] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annual review of neuroscience*, vol. 18, no. 1, pp. 193–222, 1995.

[17] Y. Sandamirskaya and G. Schöner, "An embodied account of serial order: How instabilities drive sequence generation," *Neural Networks*, vol. 23, no. 10, pp. 1164–1179, 2010.

[18] S. Schneegans and G. Schöner, "A neural mechanism for coordinate transformation predicts pre-saccadic remapping," *Biological cybernetics*, vol. 106, no. 2, p. 89–109, 2012.

[19] M. Richter, J. Lins, and G. Schöner, "A neural dynamic model of the perceptual grounding of spatial and movement relations," *Cognitive Science*, vol. 45, no. 10, p. e13045, 2021.

[20] S. Kazerounian, M. Luciw, M. Richter, and Y. Sandamirskaya, "Autonomous reinforcement of behavioral sequences in neural dynamics," in *International Joint Conference on Neural Networks (IJCNN)*, 2013.

[21] J. Tekülve, A. Fois, Y. Sandamirskaya, and G. Schöner, "Autonomous sequence generation for a neural dynamic robot: Scene perception, serial order, and object-oriented movement," *Frontiers in Neurorobotics*, vol. 13, p. 95, 2019.

[22] G. Schöner, L. Bildheim, and L. Zhang, "Toward a neural theory of goal-directed reaching movements," in *Progress in Motor Control: From Neuroscience to Patient Outcomes*, Levin, M F, Petrarca, M, Piscitelli, D, and Summa, S, Eds. Academic Press, 2024, pp. 71–102.

[23] M. Richter, Y. Sandamirskaya, and G. Schöner, "A robotic architecture for action selection and behavioral organization inspired by human cognition," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. IEEE, 2012, pp. 2457–2464.

[24] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?" *Neuron*, vol. 73, no. 3, pp. 415–434, 2012.

[25] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundations: A survey," *ACM Transactions on Applied Perception (TAP)*, vol. 7, no. 1, pp. 1–39, 2010.

[26] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 185–207, 2012.