Cognitive Science (in press, 2021)

A neural dynamic model of the perceptual grounding of spatial and movement relations

Mathis Richter*, Jonas Lins, Gregor Schöner

Abstract

How does the human brain link relational concepts to perceptual experience? For example, a speaker may say "the cup to the left of the computer" to direct the listener's attention to one of two cups on a desk. We provide a neural dynamic account for both perceptual grounding, in which relational concepts enable the attentional selection of objects in the visual array, and for the generation of descriptions of the visual array using relational concepts. In the model, activation in neural populations evolves dynamically under both the influence of inputs and of strong interaction as formalized in dynamic field theory (DFT). Relational concepts are modeled as patterns of connectivity to perceptual representations. These generalize across the visual array through active coordinate transforms that center the representation of target objects in potential reference objects. How the model perceptually grounds or generates relational descriptions is probed in 104 simulations that systematically vary the spatial and movement relations employed, the number of feature dimensions used, and the number of matching and non-matching objects. We explain how sequences of decisions emerge from the time- and state-continuous neural dynamics, how relational hypotheses are generated and either accepted or rejected, followed by the selection of new objects or the generation of new relational hypotheses. Its neural realism distinguishes the model from information processing accounts, its capacity to autonomously generate sequences of processing steps distinguishes it from deep neural network accounts. The model points toward a neural dynamic theory of higher cognition.

1 Introduction

A fundamental function of human language is to enable speaker and listener to communicate about the environment they both experience. The speaker *describes* objects or events in the environment she experiences. The listener identifies objects or events in his environment that the speaker talks about, *perceptually grounding* them. Imagine, for example, two young siblings playing with a dollhouse (Tenbrink et al., 2017). The sister describes to her brother a bed in the dollhouse, linking her perceptual experience to language. Her brother perceptually grounds that description in directing his visual attention to the bed. As a result, the siblings establish joint attention (Tomasello, 1995), enabling them to communicate about the shared environment. How may the brains of these children achieve that? And what would it mean to explain the neural processes on which that capacity is

*Affiliation of all authors: Institut für Neuroinformatik, Ruhr-Universität Bochum, 44780 Bochum, Germany. Corresponding author: Mathis Richter (mathis.richter@ini.rub.de)

based? To answer these questions we propose a neural process model that establishes the links between categorical concepts and (visual) perceptual experience.

Furthermore, after establishing joint attention, the sister may instruct her brother, for instance, to place "a nightstand to the left of the bed". In this phrase, she is referencing a concrete location in the dollhouse by using the spatial relation "to the left of". Such spatial relations may be lifted to express more abstract concepts in higher cognition (Lakoff & Johnson, 1999) (where "higher" provides itself an example of such metaphorical use of a spatial relation). Movement relations, such as "move the nightstand to the bed", are foundational for representing actions and, thus, for the majority of verbs (Pulvermüller, 2005). While simple object features, such as color and orientation, can be extracted from single locations in the visual array, relations link to both a target and a reference object. A key issue that we will address is how a neural network implementing a relation can be flexibly applied to reference objects anywhere in the visual array.

Fig. 1 illustrates key notions of the neural account of grounding and description generation of relation that we aim at in this paper. A visual sensor (bottom panel) provides input to populations of neurons that extract local features (e.g., hue values). These populations form neural maps or fields that represent visual space together with such feature dimensions (not shown). These fields are used to guide attentional selection within purely spatial fields, of which three are shown in the middle panel of Fig. 1. Positive levels of neural activation (color coded in yellow) are localized where objects have been brought into the attentional foreground. Neural nodes illustrated in the top panel reflect activity in small populations of neurons that represent concepts such as RED and ABOVE. Language-like propositions, such as "the red object above the green object" are then represented by activation patterns within these nodes. Reciprocal connectivity between nodes and neural fields define the perceptual meaning of the concepts. How relational concepts must be connected to the fields that perceptually ground them is a key topic of this paper.

In *perceptual grounding* (orange arrows on the left of Fig. 1), a neural representation of a phrase guides a process of visual search to direct attention to an object that matches the relation. In the figure, the phrase "red above green" is encoded by an activated "red" concept node for the target (top left), an activated "above" concept node for the relation (top center), and an activated "green" concept node for the reference object (top right). Grounding is achieved when in the activation field representing target locations (middle left) an activation peak is positioned over the location of the red object that is above the green object. The activation field representing reference object locations (middle right) has a peak at the location of that green object. The activation pattern in the relational field (middle center) reflects the same target location now centered on the location of the reference object (the areas marked by white ellipses correspond).

In *description generation* (blue arrows on the right of Fig. 1), the attentional selection of objects in the scene is based on their salience and their match to spatial or movement relations. In the figure, the red object on the right of the scene is brought into attention by salience. This leads to the selection of a matching reference object and relation which drives the activation of a neural representation of the phrase "red above green" in the concept nodes.

In our conception, a neural process account does not necessarily entail a detailed mapping of neural populations on particular brain areas or sub-networks. Such a mapping may ultimately be achievable, but that is not the goal we set ourselves. Instead, we constrain the model by three principles that we believe broadly define an embodied, neural

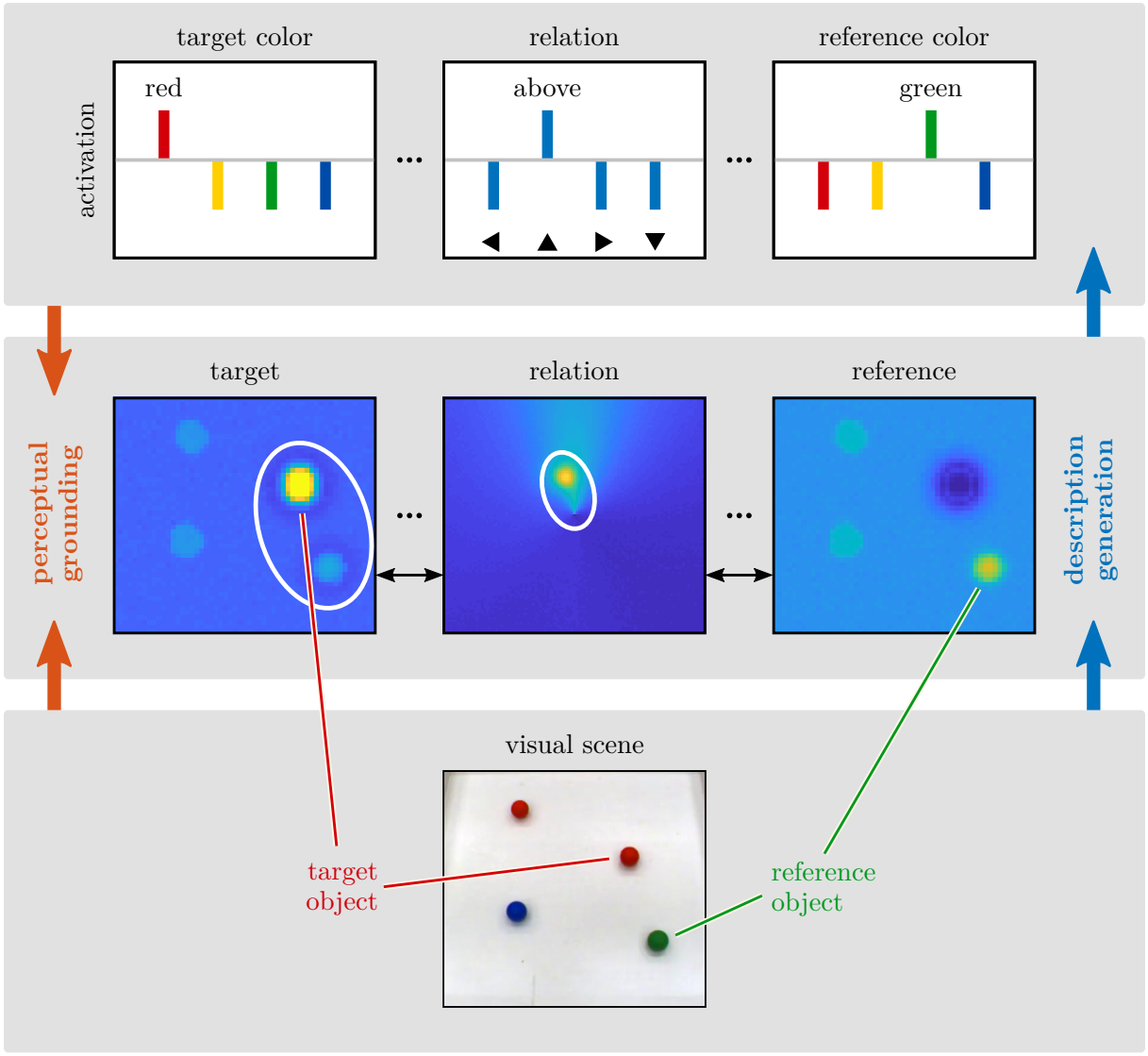

Fig. 1: The two directions of linking language to visual cognition: perceptual grounding (orange arrows) and description generation (blue arrows). The language phrase that the model is representing here is "the red object above the green object". See text for description.

dynamic perspective on cognitive modeling.

First, *the function of a neural network is determined by its connectivity*, the fundamental postulate of connectionism (McClelland et al., 2010). Activation patterns in neural networks arise in response to inputs or are self-generated within the network based on recurrent connectivity. Neural connectivity is fixed on the short time scale of a single instance of a mental or motor act, but may change through adaptation and learning across multiple instances of experience.

Second, neural processes are *embedded in the body and connected to its sensors and motor systems* (Barsalou, 1999; Clark, 1999; M. Wilson, 2002). Neural mappings from the sensory surfaces to the brain reflect the continuous ways in which objects, scenes, or events may vary (e.g., as objects vary in pose, shape, and surface properties). Neural mappings from the brain to the motor surfaces reflect the continuous ways in which movement behaviors may vary (e.g., in direction, speed, or movement time). The similarity

of patterns of connectivity in primary sensory and motor areas (i.e., locally excitatory and globally inhibitory) to patterns of connectivity in areas of the brain further removed from the sensory and motor surfaces supports the strong embodiment hypothesis according to which higher cognitive processes share properties with sensory-motor processes, in particular, sensitivity to continuous dimensions that ultimately reflect sensory-motor properties (Gärdenfors, 2014). Higher cognitive processes may couple to sensory-motor processes, at least intermittently. This implies that they must have stability properties, so that they may track time-varying input and resist disturbances of motor states.

Third, and extending the embodiment hypothesis, neural activation evolves in *continuous time* as described by neural dynamics. Decisions, sequences of thoughts or actions emerge at discrete times from such underlying continuous time. How this happens is in need of neural explanation (and will be addressed here). This principle contrasts neural dynamics to computational models that account for discrete events based on digital computer notions of a central clock or updating cycle, or based on information processing notions such as "firing a production rule".

By building mathematical models that may actually generate the outcomes of perceptual grounding and of description generation from visual input, we move beyond verbal or conceptual accounts of grounding that are broadly consistent with some of these principles (e.g., Barsalou, 1999; Gibbs & Colston, 1995; Langacker, 1986; Talmy, 1988). We also move beyond computational accounts (e.g., Pastra & Aloimonos, 2012; Roy, 2008) that describe these processes at an abstract level. For example, the relation "the nightstand to the left of the bed" may be modelled as a function or operator, LEFT(NIGHTSTAND, BED), that takes two arguments, the positions of the nightstand and the bed, and then returns a truth value or a graded measure of certainty. Such a computational view of perceptual grounding is incompatible with the principles of neural processing stated above. It does, however, achieve a powerful flexibility by enabling arbitrary arguments to be passed to the function. In fact, providing an account consistent with our neural principles for that computational flexibility is a fundamental challenge: How may neural connectivity that instantiates a particular cognitive operation be brought to bear on a wide range of possible "arguments"? That connectivity can only receive "arguments" to the extent to which it is connected to the neural sub-populations from which these "arguments" arise. Many convolutional (deep) neural networks sidestep this issue by weight sharing, instantiating the connectivity of an operator invariantly across space. We address this issue by showing how an invariant representation of space can be obtained, on which a single neural instantiation of an operator can be applied across all possible arguments. The key idea is that the neural representation of the visual array is actively coordinate transformed to become centered on a potential reference object (center panel of Fig. 1).

The model is based on dynamic field theory, a mathematical and conceptual framework for neural process accounts consistent with the three stated neural principles. The relevant concepts of dynamic field theory are briefly reviewed next, followed by a description of the model and a series of simulations that demonstrate how the model generates the time courses of neural activation that bring about perceptual grounding and description generation in the sense suggested in Fig. 1.

2 Dynamic field theory (DFT)

Dynamic field theory (Schöner et al., 2016) elaborates and formalizes the three principles of embodied neural dynamic process accounts. DFT extrapolates evidence that mental

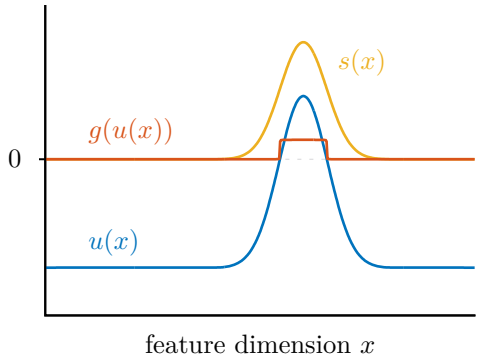

Fig. 2: A dynamic neural field of activation, $u(x)$ (blue), is defined over a single feature dimension x by virtue of localized input, $s(x)$ (yellow). The field produces output only where its activation exceeds the threshold at 0 of a sigmoid function $g(u(x))$ (red).

and behavioral states are best explained in terms of the activity in small populations of neurons (Cohen & Newsome, 2009; Panzeri et al., 2015; Wu et al., 2011). In that view, the population level is privileged to account for the neural processes on which thinking and acting is based (Schöner, 2019). The building blocks of DFT models are thus patterns of activation in populations of neurons (illustrated in Fig. 2). These patterns are characterized by graded activation variables, u (blue line in Fig. 2), that reflect how close a neural population is to affecting down-stream neural populations onto which it projects. Projection onto other neural populations is governed by a sigmoidal threshold function, $g(u)$, (red line) so that only activation larger than the threshold at zero is transmitted. In the brain, activation is determined by the membrane potentials of neurons within a population and the threshold function is established by the spiking mechanism. The mesoscopic level of description of DFT does not take into account these neuro-physical mechanisms in what can be derived under some conditions as the mean-field approximation (Faugeras et al., 2009).

Activation patterns of a neural population may span continuous, low-dimensional feature spaces, x, by virtue of forward connectivity from sensory surfaces to the population or from the population to motor systems, both characterized by tuning curves. The resulting neural maps can be viewed as *dynamic neural fields*, $u(x)$, defined over *continuous feature spaces*, x, when we assume that the discrete sampling of the feature spaces by individual neurons is not functionally relevant (Erlhagen et al., 1999). Consequently, the units of representation are not individual neurons but peaks of activation localized along the feature dimension (as illustrated in Fig. 2) that pass their supra-threshold activation on to down-stream populations in ways that may reflect the position of the peak within the field. Neural populations that do not span continuous feature spaces are modelled by *dynamic neural nodes* that represent categorical concepts. This often happens in the form of sets of such nodes that are inhibitorily coupled to enable "winner-take-all" selection.

The activation patterns, $u(x,t)$, of dynamic neural fields and nodes evolve continuously in time, t, as described by a *neural dynamics*,

$$\tau \dot{u}(x,t) = -u(x,t) + h + s(x,t) + \int dx'\, k(x-x')\, g(u(x',t)). \tag{1}$$

This mathematical formalization, which dates back to the 1970s (Amari, 1977; H. R. Wilson & Cowan, 1973), lifts the dynamics of neural membranes to the population level. The rate of change, $\dot{u}(x,t)$, of the activation pattern depends on a time scale parameter, τ, that

determines how quickly the dynamics converges, a resting level, $h < 0$, that determines the level to which activation converges without external input, localized external inputs, $s(x,t)$, from other dynamic neural fields or sensors, and homogeneous patterns of within-population connectivity, $k(x - x')$, explained below.

Stability, the capacity to resist change, is central to the embodied perspective (Spencer & Schöner, 2003) and enables the coupling of neural dynamic processes to sensory-motor systems. Localized peaks of activation of dynamic neural fields (Fig. 2) are *attractor states* of the neural dynamics, stabilized by the strong recurrent connectivity within the field. That connectivity, expressed here as a function, $k(x - x')$, of the distance between two points in the field, is positive over short distances, stabilizing peaks against decay. It is negative over longer distances, stabilizing peaks against distractor input. This pattern of connectivity is found in the brain (Douglas & Martin, 2004; Jancke et al., 1999) and is commonly invoked in models of cortical function (Rutishauser et al., 2010). The integral term in the equation formalizes that this same connectivity pattern is applied to any location with supra-threshold activation, $g(u(x'))$, and is summed along the entire feature dimension, x'.

The sub-threshold state, $u(x,t) = h + s(x,t)$, is also an attractor solution for small and slowly varying inputs $s(x,t) < -h$. This attractor solution disappears in the *detection instability* when increasing input pushes activation above zero and induces a switch to a peak solution. This happens at a discrete moment in time even if input increases continuously in time. The detection instability thus explains how discrete time events emerge from continuous time neural dynamics.

Fields support *selection*, in which a peak forms over one among multiple local maxima of input. Once a local maximum has been selected, the peak tracks any changes in input. Fields may also support multiple localized peaks of activation, depending on the strength and spatial dependence of inhibitory coupling. Activation peaks may be sustained by the intra-field coupling after inducing localized input has been removed, providing an account for working memory of feature values (Durstewitz et al., 2000; Johnson et al., 2009). The dynamic stability of attractor states provides structural stability (Perko, 2001), in which attractors and their instabilities—dynamic regimes—persist even as the dynamic equation is gradually changed. Dynamic regimes remain invariant, therefore, under graded change of parameters and inputs, but also when fields are coupled to other fields or nodes. Neural dynamic architectures may thus be characterized by the dynamic regimes of their component fields, a form of modularity.

When neural dynamic fields depend on multiple different feature dimensions, they form bound representations of those features. For instance, a field depending on both visual space and on color binds these two features (see the color/space perception field in Fig. 4 for an example). Projections from such bound representations provide a neural instantiation of functions of the bound feature dimensions.

An important class of such functions are active coordinate transforms, which can be viewed as steerable maps (Fig. 3; Pouget & Sejnowski, 1997; Schneegans & Schöner, 2012). The figure illustrates how target objects may be transformed into a coordinate frame centered on a reference object. The *transformation field* binds the spatial positions of target and reference objects. Sub-threshold ridge input from the target and reference field generates peaks at locations that combine the respective spatial locations (marked in bright/yellow shading). Summing along the diagonal of the transformation field and projecting onto the *relational field* as sketched generates a spatial representation of the target objects that is centered on the position of the reference object. Implementing

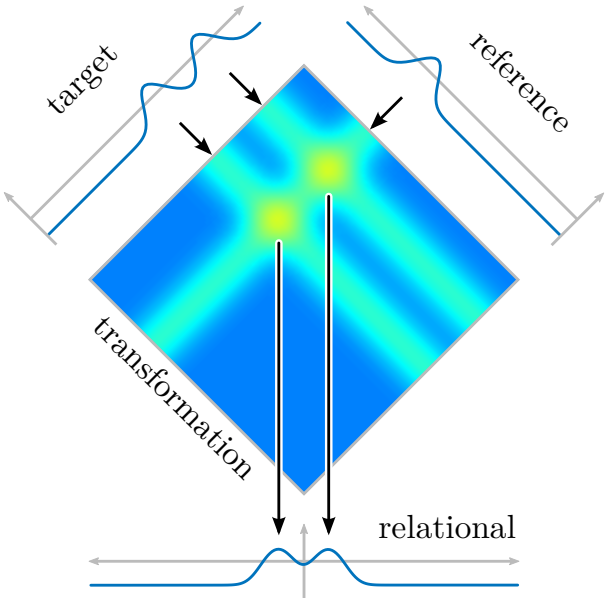

Fig. 3: A steerable neural map implemented as as two-dimensional transformation field. Activation is color coded, with yellow/bright shading marking supra-threshold, blue/dark marking sub-threshold levels of activation. Here, the reference object is positioned between the two target objects, whose representation in the relational field is thus symmetrical around the center.

spatial relations as connectivity patterns of the *relational field*, we exploit such coordinate transforms to apply these connectivity patterns invariantly across different locations of the reference object.

3 Model

The architecture we describe in this section is a neural process model of perceptual grounding and description generation. When presented with video input of a visual scene, as well as a phrase about an object in that scene, it is able to perceptually ground that phrase by bringing the designated object into the attentional foreground. When presented with video input alone, it is able to generate a phrase describing any object in the scene.

We organize the description of the model in four parts, as illustrated in Fig. 4. It is important to keep in mind, however, that the model is essentially one big dynamical system described by a large set of coupled integro-differential equations (Richter, 2018). Each part is simply a set of activation variables evolving in time, and the functional interpretation of activation states is primarily a mental guide to us, the modeler and the reader, to keep track of what is going on in the model.

3.1 Perception and visual search

This part of the model is a simplified version of a more comprehensive neural dynamic model of visual search (Grieben et al., 2020) that contains the perceptual front-end. Visual search attentionally selects objects in the scene based on two pathways: bottom-up inputs that arise from the sensory surface guide attention to salient objects, while top-down inputs from language may guide attention toward objects that have particular features.

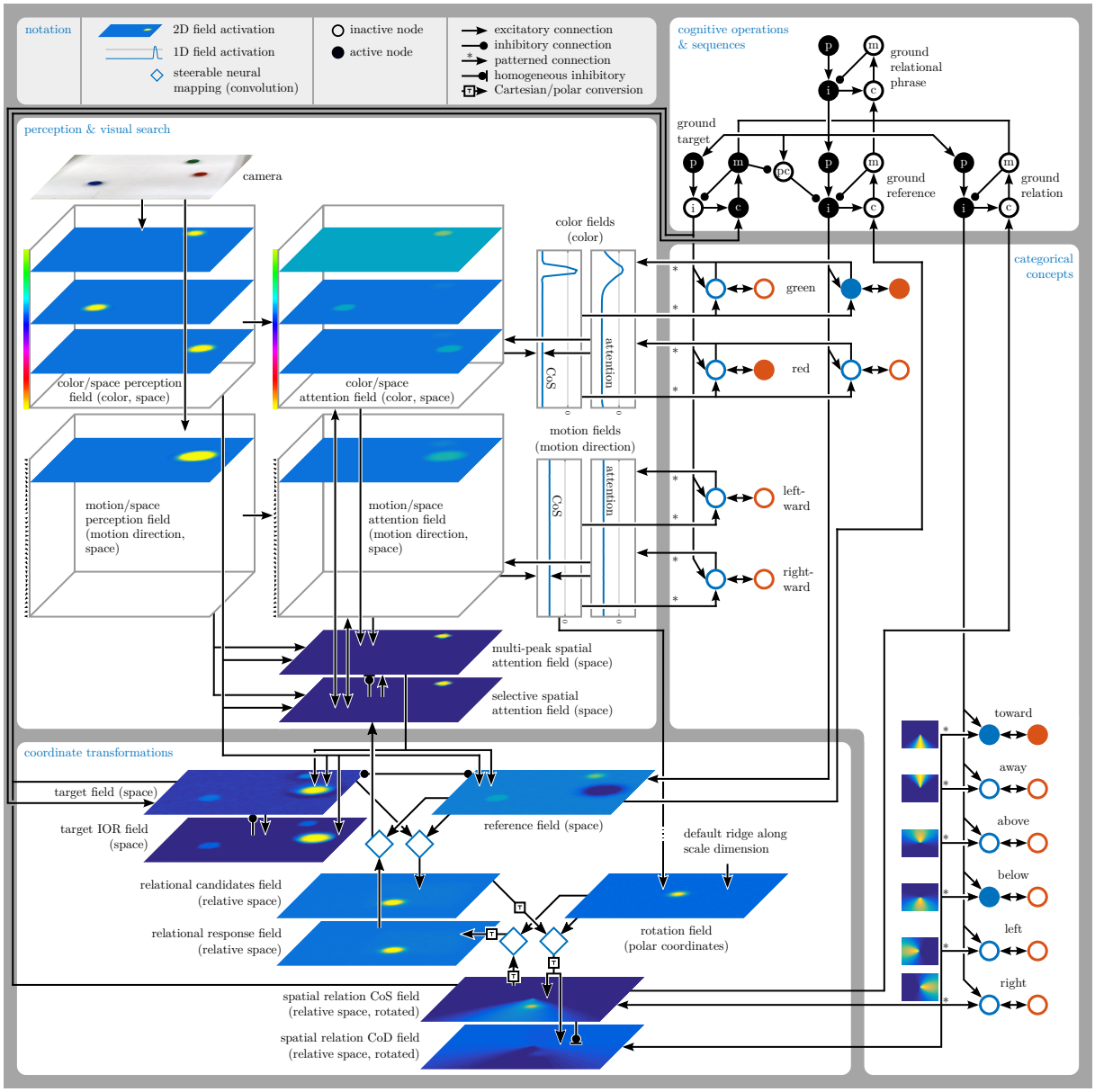

Fig. 4: Schematic overview of the model showing an activation snapshot during the grounding of the phrase "the red object moving toward the green object". White regions highlight four parts of the model that we organize our description around. For three-dimensional fields, two-dimensional slices of activation are shown. The connectivity supporting categorical concepts and sequence generation is not shown in full detail.

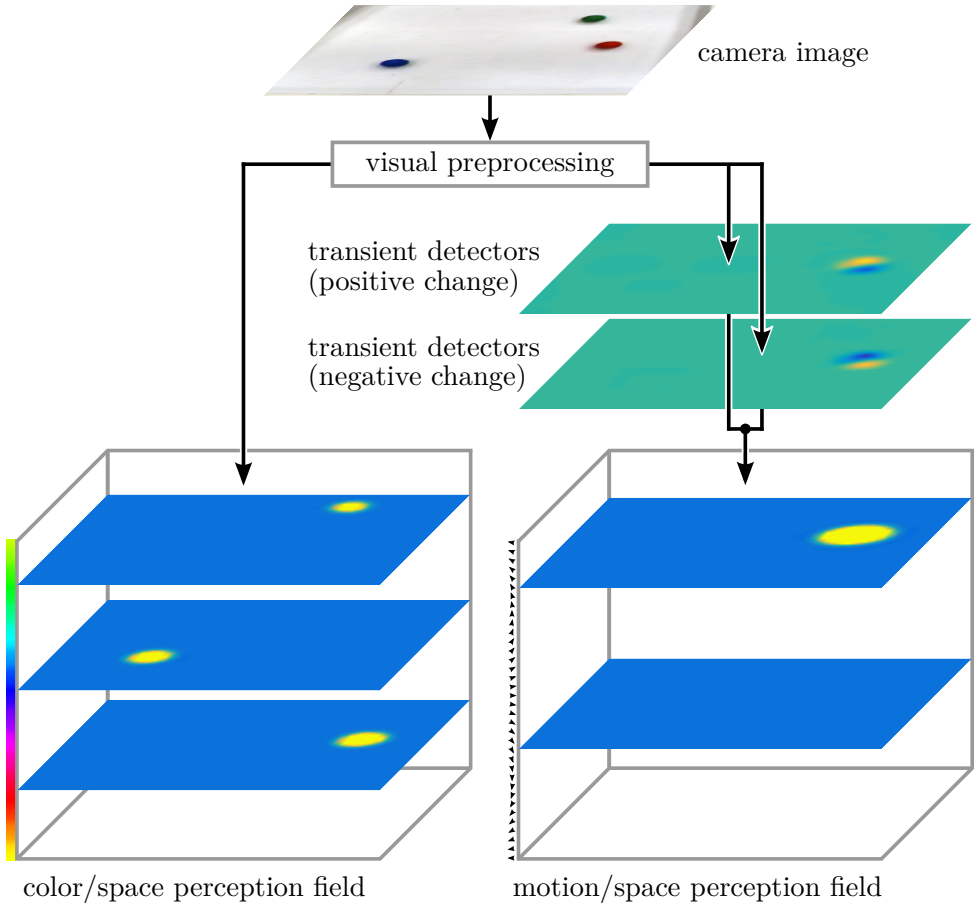

Fig. 5: Schematic overview of the perceptual front-end. For three-dimensional fields, two-dimensional slices of activation are shown.

Perceptual input to the model comes from a camera or video stream. Each frame goes through several preprocessing steps, implemented algorithmically here as a shortcut for neurally plausible models of early human visual processing (Lomp et al., 2017). During preprocessing, input to the model is scaled to the color saturation of visible objects, so that high levels of input arise near objects with uniform, saturated colors, while low levels arise from locations in the visual array with colors of low saturation (e.g., black and white).

The model builds a neural representation of all visible objects as an activation pattern in two three-dimensional dynamic neural fields (top left white box in Fig. 4). The *color/space perception field* receives the preprocessed video data as input, and is defined over the two spatial dimensions of the camera image and over hue (color), building localized peaks in response to mono-chrome objects (see also Fig. 5). The *motion/space perception field* is defined over the same two spatial dimensions of the camera image and over the feature dimension of motion direction. Moving objects induce peaks of activation in this field, while stationary objects do not. The motion signal is extracted based on the counter-change model of motion detection (Berger et al., 2012). Motion is detected from a location at which image saturation changes toward back-ground to a location at which image saturation changes away from back-ground. These changes are signaled by two sets of transient detectors (Fig. 5).

At the core of the top-down pathway of visual search are two three-dimensional dynamic neural fields, the *color/space attention field* and the *motion/space attention field*, defined over the same dimensions as the two respective perceptual fields. A peak in these attention

fields represents the attentional selection of the object at the corresponding location and feature value. Each field receives input from its corresponding perceptual field, which by itself is not sufficient to induce peaks, leading merely to subthreshold bumps of activation. Only with additional "top-down" input that overlaps with these bumps is the detection instability reached and a peak formed.

Attention to a particular color is modeled by input from the *color attention field*, defined over hue. The projection from a one-dimensional to a three-dimensional field can be visualized as a sheet of input, that pushes activation up in all spatial locations at the hue value represented in the color attention field. For instance, in Fig. 4, this input for green pushes locations of green objects through the detection threshold. Attention to a particular motion direction is modeled analogously by input from the *motion attention field*, defined over motion direction.

Spatial attention is modeled by input from the *selective spatial attention field*, defined over the spatial dimensions of the camera image, and operated in a dynamic regime in which only a single peak may form at a time. Spatially localized activation can be visualized as a cylinder oriented along the feature dimension (color or movement direction) and promotes the selection of objects at the position of the cylinder. Coupling between the selective spatial attention field and the three-dimensional attention fields is bidirectional along the shared spatial dimensions. Note that the different features of each object are represented in separate neural populations (separate three-dimensional fields), leading to the famous binding problem. The bidirectional coupling solves the binding problem in the manner of Feature Integration Theory (Treisman & Gelade, 1980) by linking different features of each object through the spatial dimensions shared by all feature-space fields (Grieben et al., 2020; Schneegans, Spencer, et al., 2016).

Input from the perception fields to the selective spatial attention field implements a simple bottom-up salience mechanism for spatial attention in the absence of top-down cues. Salience increases with the visual size of an object, the saturation of its colors, and when the object moves.

Multiple object positions may be simultaneous relayed to other fields by the *multi-peak spatial attention field*, which receives the same input as the selective spatial attention field but operates in the dynamic regime that allows multiple peaks to form at the same time. The two fields together form a functional unit whose spatial selectivity can be modulated by homogeneous input to the selective spatial attention field. A peak in that field inhibits all other peaks in the multi-peak spatial attention field, effectively switching the spatial attention system to a selective mode. This flexibility enables the system to either attentionally select one out of multiple candidate objects based on their object features and saliency alone or to defer that attentional selection decision to other fields, where it may be based on the match of the objects with a relation.

3.2 Categorical concepts

The top-down input guiding visual search ultimately comes from a neural representation of a phrase. In the model, the interface to language is defined by categorical concepts of color, motion direction, spatial relations, and movement relations (bottom right box in Fig. 4), mapping words to concepts (Jackendoff, 2012). We do not model language processing itself, assuming that the output of such processing activates *memory nodes* that represent the phrase to be grounded (e.g., RED, RIGHTWARD). Conversely, the model generates a scene description by activating the appropriate memory nodes.

Each memory node is coupled excitatorily and reciprocally to an associated *production node*, which may become activated and thereby have an impact on the rest of the model. The connectivity pattern between a production node and a dynamic neural field, defined over color, motion direction, or space, instantiates the perceptual meaning of a concept. For instance, Fig. 4 shows how the production node of the color concept GREEN activates a region of the color attention field, in which hues of green are represented. The pattern of connectivity between the node and the field is modeled as a Gaussian function along the color dimension, centered on green. Activating the node may ultimately lead to the attentional selection of green objects. Conversely, the GREEN concept node becomes activated when a green object is in the focus of spatial attention.

In language and thought, we assign roles to the objects we talk about (Landau & Jackendoff, 1993) such as when the red object in "the red object to the left of the green object" is the target of the spatial relation and the green object is the reference object of that relation. The model has separate neural representations of object attributes and of their perceptual grounding for every role an object may appear in. So, every color concept is represented by pairs of memory and production nodes for the roles as target and as reference. The color concept production nodes connect to the same color attention field with the same connectivity in the two roles. For example, the production nodes representing GREEN as a target and GREEN as a reference have the same pattern of connectivity to the color attention field. One may think of the nodes for the two roles as modeling two sub-populations that share the same overall connectivity pattern toward the perceptual system but are wired up differently toward the conceptual system. In fact, roles themselves do not have any perceptual meaning; what makes a neural representation play a specific role is determined only by its connectivity to and from other parts of the architecture.

The color concepts, RED, YELLOW, GREEN, and BLUE, and the concepts of motion direction LEFTWARD, RIGHTWARD, UPWARD, and DOWNWARD, are realized in the model to provide attributes of objects that guide attentional selection. Their pattern of connectivity is modeled as a Gaussian over their respective feature dimensions, analogously to what is outlined for the color concept GREEN above.

The concepts of spatial relations TO THE LEFT OF, TO THE RIGHT OF, ABOVE, BELOW, and of movement relations, TOWARD, and AWAY FROM, are realized in the model to provide relations between pairs of objects. If motion is perceived in the scene, the model activates concepts of movement relations; for static scenes, concepts of spatial relations are activated. The connectivity pattern of these relational concepts is modeled (Lipinski et al., 2012) to fit behavioral rating data (Logan & Sadler, 1996). Connectivity is excitatory in the appropriate spatial region relative to the center of the spatial field (patterns are shown at the bottom right of Fig. 4). The negation of the concept (e.g., NOT LEFT) is represented by a connectivity pattern, in which the spatial region is inhibited rather than excited.

Based on these connectivity patterns, representations within the model may match a categorical concept to varying degrees. A match decision is made for sufficient overlap of a bump of activation with the connectivity pattern, fulfilling what is called the condition of satisfaction (CoS) (Searle, 1980); a non-match decision is made for sufficient overlap of a bump with a connectivity pattern of inverted polarity, fulfilling the condition of dissatisfaction (CoD). Fig. 4 shows an example in the spatial relation CoS field (bottom center), where a localized bump of activation close to the center of the field overlaps with the connectivity pattern of the movement relation TOWARD.

3.3 Coordinate transformations

The fixed connectivity patterns of relational concepts can be brought to bear on any location of the visual array by actively transforming the neural representation of target objects into a coordinate frame centered on the location of reference objects (Fig. 3). The transform is based on a bound representation of target and reference objects, which first requires neural representations dedicated to objects in these two roles. Their spatial positions are represented in the *target field* for targets and in the *reference field* for reference objects. The two fields inhibit each other such that any given spatial location is only ever represented in one of the two fields. Both fields receive input from the multi-peak spatial attention field, reflecting the spatial locations of objects that are currently in the focus of attention. Processing is organized sequentially (see Section 3.4) to direct attention first to target, then to reference objects.[1] At each step, the corresponding spatial attention field is boosted by a homogeneous input. When the reference field is boosted, the selective spatial attention field is deactivated, enabling multiple candidate positions to be represented in the reference field. When the target field is boosted, the selective spatial attention field is activated, enabling a selection decision for a single target object. Both the target field and the reference field receive additional input from the perception fields so that their peak solutions, once instantiated, may track moving objects.

Given the two-dimensional spatial representation of objects in the model, the transformation field of the active coordinate transform is four-dimensional. To reduce the computational load, we approximate the transformation fields using algorithmic convolution and correlation functions. The model uses coordinate transformations twice in succession (bottom left white box in Fig. 4). First, to actively transform the neural representation of target locations into a reference frame that is centered on a reference object, the reference field steers how the target field projects onto the *relational candidates field*. In the example of Fig. 4 (top right transformation, depicted by a blue diamond), the relational candidates field represents the position of the red target object relative to the position of the green reference object. The second transformation rotates the resulting representation so that it is centered on the motion direction of the target object. The *rotation field* represents the movement direction of the target object and steers that transformation (Fig. 4, bottom right diamond). The outcome is projected onto two fields, the *spatial relation CoS (condition of satisfaction) field* and the *spatial relation CoD (condition of dissatisfaction) field*.

All relational concepts, such as TOWARD and TO THE LEFT OF, are instantiated by connectivity patterns between the associated nodes and these two fields (see Section 3.2). Positions of the target and reference object that match the relation induce a peak in the spatial relation CoS field. In the absence of a match, positions of the target and reference object at non-matching locations induce a peak in the spatial relation CoD field. This signals a failure to find a target and leads to a repeated search (see below).

Once a matching object has been selected in the spatial relation CoS field, its peak is projected back into the selective spatial attention field. This requires transformations (Fig. 4, diamonds on the left) that invert those described above as well as an intermediate representation in the *relational response field*. The output of these transformations brings the attentional focus to the spatial position of the selected reference object.

[1] This order can be made flexible, to reflect, for instance, the order of speech input. It is fixed here to enable a simple form of generating descriptions, in which the most salient object in the scene is attended first and becomes the target object.

3.4 Cognitive operations

As the model solves its tasks, it goes through a set of *cognitive operations* by transitioning between qualitatively different neural states that are demarcated by dynamic instabilities (see Section 2). Cognitive operations typically entail that a peak is generated or destroyed, for example to bring objects of a certain color into the attentional foreground, to make a selection of one object from multiple candidates, to create a working-memory representation of an object, or to delete such a representation. Fundamentally, all processing in the model runs in parallel, so that transitions may happen asynchronously in different parts of the model. In some cases, operations must be performed in sequence due to constraints to operate on one object at a time. For instance, to ground relations, the target and the reference object must be identified and their locations be entered into the target and reference field. This requires sequential attentional selection of candidates for the two roles.

The coordination of such neural operations is based on structured sub-networks of four nodes (Richter et al., 2012; Sandamirskaya & Schöner, 2010). The complete network that organizes perceptual grounding and description generation consists of 21 such sub-networks and is too complex to be fully illustrated here; we only sketch parts of it in Fig. 4 (top right). The sub-networks are organized in a four-level hierarchy. The highest level controls which of three different tasks the model performs: grounding a single object, grounding a relation, and generating a description of an object or a relation. The second level controls the grounding of the three elements of a spatial phrase: the target and reference objects, and the spatial relation. Two sub-networks control the inhibition of different parts of the architecture to "clean" the fields used for grounding at appropriate points in a sequence. The third and fourth levels control more detailed aspects of the grounding operation such as bringing individual fields into dynamic regimes in which they can form peaks. Different sub-networks at the lower levels are activated in varying sequential orders by sub-networks at higher levels, which leads, overall, to a variety of qualitatively different behaviors of the model.

To explain the structure and use of the sub-networks that coordinate neural operations, we follow along the example illustrated in Fig. 6. In the example, the cognitive operation of grounding a relational phrase is broken down into three cognitive operations at a lower hierarchical level: grounding the target object, the reference object, and the spatial relation. Each of these three cognitive operations is represented and coordinated by a dedicated sub-network of four nodes each.

The *intention node*[2] (labeled "i" in Fig. 6) is the output unit of this small network that projects onto other sub-populations or sub-networks which bring about a particular cognitive operation. When its activation exceeds the threshold of the sigmoidal function, it modulates the dynamic mode of its target networks. For instance, the intention node of the operation "ground target" enables the target field to build peaks by pushing the field closer to the detection instability. When additional localized input from other sources projects onto a set of fields, including the target field, only the target field may generate activation peaks, while the others may not. In this way, intention nodes may effectively redirect the flow of coupling within a neural dynamic architecture. The *condition of satisfaction (CoS) node* (labeled "c") is pre-activated by the intention node, and receives additional input from other subnetworks that signals that the cognitive operation enabled by the intention

[2] *Intentions* are neural states that are *about* the world. A match between the contents of an intentional state and a state of the world is the intention's condition of satisfaction (CoS) (Searle, 1980).

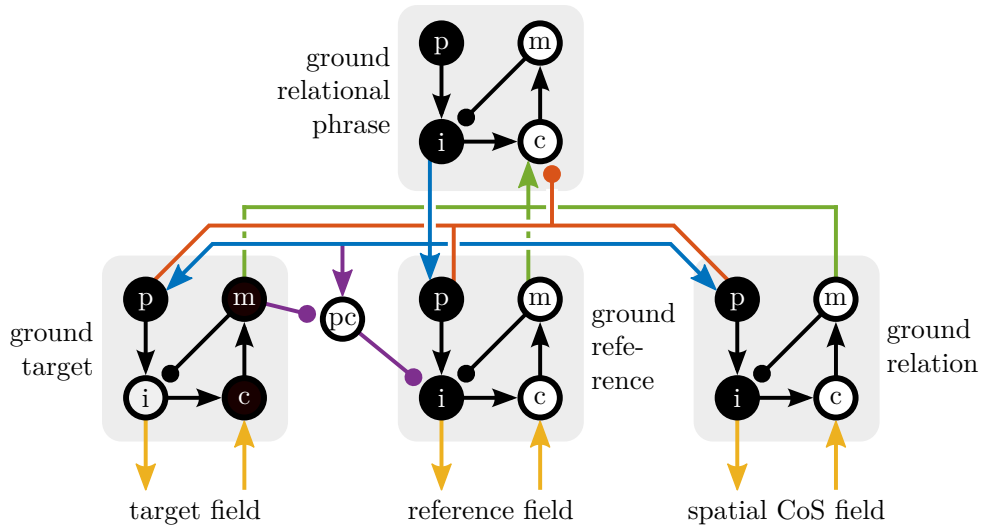

Fig. 6: An exemplary hierarchy of sub-networks (gray boxes) that organize cognitive operations in the model. Circles denote neural nodes ("p": prior intention, "m": condition of satisfaction (CoS) memory, "i": intention, "c": CoS, "pc": precondition). Filled black circles denote active nodes, white circles denote inactive nodes. All nodes are self-excitatory, not graphically represented here. Regular arrows mark excitatory connections; lines ending in a small filled circle mark inhibitory connections. Different line colors are for visual aid only. See text for details.

node has led to the expected outcome. For the cognitive operation "ground target", such an input would arise when a peak has formed in the target field, signaling the successful completion of the operation and activating the CoS node. This state is propagated to the *CoS memory node* (labeled "m"), which inhibits the intention node, causing it to become deactivated. This, in turn, removes input to the CoS node, turning that node off, and thus effectively deactivating the entire sub-network. A memory of its successful completion is kept in the CoS memory node, however, which remains activated based on self-excitatory coupling.

Constraints on the serial order in which cognitive operations are performed are imposed by the coupling structure of the *precondition node* (Richter et al., 2012). Fig. 6 illustrates one instance of such a precondition node (labeled "pc", purple lines), which ensures that the operation "ground target" successfully grounds the target object before the operation "ground reference" begins grounding the reference object. An activated precondition node inhibits the intention node onto which it projects. That intention node may, therefore, become activated only once a preceding cognitive operation, the precondition, has terminated successfully. Its associated CoS nodes inhibits the precondition node and releases the intention node from inhibition.

The mechanism by which precondition nodes enforce sequentiality is akin to chaining (Henson & Burgess, 1997), as one cognitive operation enables the activation of its successor as soon as the first has terminated. A challenge for chaining theories is generating different sequences that go through the same element. To enable such "re-use" of elements, the sequential activation of cognitive operations may be organized hierarchically as illustrated in the example in Fig. 6. On the lowest hierarchical level, the intention nodes project directly onto the fields of the model. Their CoS nodes receive input from those fields signaling that an operation has been completed. A particular serial order at that lower

level may be imposed by precondition nodes. Intention nodes at a higher level may project onto lower level *prior intention nodes* (labeled "p") and CoS memory nodes (illustrated by blue lines in Fig. 6). The projection is sufficiently strong to activate the prior intention nodes but not the CoS memory nodes. The activated prior intention nodes, in turn, inhibit the CoS node of the higher level network (orange lines in Fig. 4). A higher level CoS node may thus become activated only when excitatory input from the CoS memory nodes of the lower level operations becomes activated (green lines). The CoS of higher level operations consists, therefore, of the successful completion of all lower level cognitive operations. Through that pattern of connectivity, multiple higher-level operations may recruit overlapping subsets of lower-level operations while enforcing different sequential orders through the activation of appropriate precondition nodes.

As a sequence of cognitive operations unfolds, the results of the perceptual grounding or description operations are activation peaks in the corresponding target or reference fields or activation states in concept nodes. These activation patterns carry state from one sequential operation to the next. This is prominent in the *target IOR field*, which holds a representation of all target objects that have been tried in a grounding sweep, analogous to the notion of *inhibition-of-return (IOR)* in accounts of visual attention (Itti & Koch, 2001). By projecting inhibitorily to both the selective spatial attention field and the multi-peak spatial attention field, this intermediate representation biases attentional selection toward objects that have not been previously examined.

4 Results

We evaluated the performance of the model in numerical simulations that made use of *CEDAR*,[3] an open-source software framework in which DFT models can be composed graphically and their parameters can be adjusted interactively (Lomp et al., 2016). We performed a total of 104 simulations. Of these, 89 involved *grounding tasks* in which the model is presented with a visual situation and must attentionally select the object described by a relational phrase. The remaining 15 simulations were *description tasks* in which the model must generate a relational phrase by observing a visual situation.

The visual situations presented to the model varied with respect to the number of objects in the scene, the number of moving objects, the number of objects that match a description, the number of distractor objects, and whether or not distractor objects matched at least some features relevant to the task. Each combination of these factors was captured by at least one visual situation from a data set of 82 videos.[4]

Grounding tasks were performed by activating the memory nodes that represent the concepts in the phrase to be grounded and supplying a visual stimulus. For description tasks, only the visual stimulus was supplied. All tasks were then initiated by activating a task node at the top level of the control system (all through a user interface in CEDAR). Once activated, the model acted autonomously, without further intervention.

To evaluate the performance of the model qualitatively, we observed the neural activation patterns evolve in the simulator and determined if the model generated the expected activation patterns. These depend on the different conditions, but typically entail generating an activation peak within the target and reference fields at the correct location for grounding tasks, and, in addition, generating the correct pattern of activation in the

[3]https://cedar.ini.rub.de

[4]The video data set is publicly available at https://osf.io/emq3n/files under "dataset".

concept nodes for description tasks. We checked that the CoS of all relevant cognitive operations were fulfilled.

The overall outcome is that the model generates the expected activation patterns in all 104 simulations for a single set of model parameters[5] and within a single sweep of each video. For all grounding tasks, the model grounds the given phrase in the scene whenever that is possible; for all description tasks, it generates a phrase describing the given scene, where possible.

Below we describe how the model performs the different tasks by illustrating the time courses of relevant parts of the model for exemplary cases. The simulations not shown are of the same general nature and can be illustrated in the same way (Richter, 2018).

4.1 Grounding single feature attributes

To assess the feature-based attention and bottom-up saliency pathways of the model, we evaluated grounding tasks in which a single feature value is provided as an attribute of a target object, a simple form of visual search. A representative example of this group, in which the model grounds the phrase "the red object", is illustrated in Fig. 7. At time t_1, there are peaks in the color/space perception field representing the spatial position and color feature of the four objects in the scene. The peaks produce localized subthreshold activation patterns in the color attention field. In Fig. 7 the activation of this three-dimensional field is illustrated in two two-dimensional projections, in each case marginalizing the third dimension by taking the maximum along that dimension. This leads to color vs. horizontal space in the fifth row, and vertical vs. horizontal space representations in the the sixth row.

Between times t_1 and t_2, the phrase "the red object" is encoded by activating the concept memory node for RED (through CEDAR). The grounding processes are initiated by giving input to the top level task node that encodes grounding of a single target object, which activates the intention node for searching for a target object (top panel).

At time t_2, the boost from the target intention node activates the production node for RED (transparent red bar, third row). Through its projection to the color attention field, the target color production node induces a peak there centered on red (fourth row). This peak projects into the color/space attention field centered at red. In Fig. 7, this is visible as a line along the horizontal spatial dimension (fifth row) and overall higher activation along both spatial dimensions (sixth row). Input from the red object in the scene overlaps with that projected activation.

Reciprocal interaction between the color/space attention field and the multi-peak spatial attention field and selective spatial attention field (not shown), enhances the activation at this location, visible at time t_3 as a vertical line of input in the color/space attention field (fifth row). From this loop of interaction, a peak in the color/space attention field emerges at the locations that matches the top-down feature cue, red. It brings about the attentional selection of the red object, visible in the target field at this time: The phrase "the red object" has been perceptually grounded.

Because the CoS of the various intention nodes have been met, activation begins to decay. By time t_4, most of the activation in the model has returned below threshold. What remains active are the target color memory node representing the original phrase together with its perceptual grounding in the target field.

[5]The parameter set is publicly available as part of a human-readable configuration file for CEDAR at https://osf.io/emq3n/files under "config".

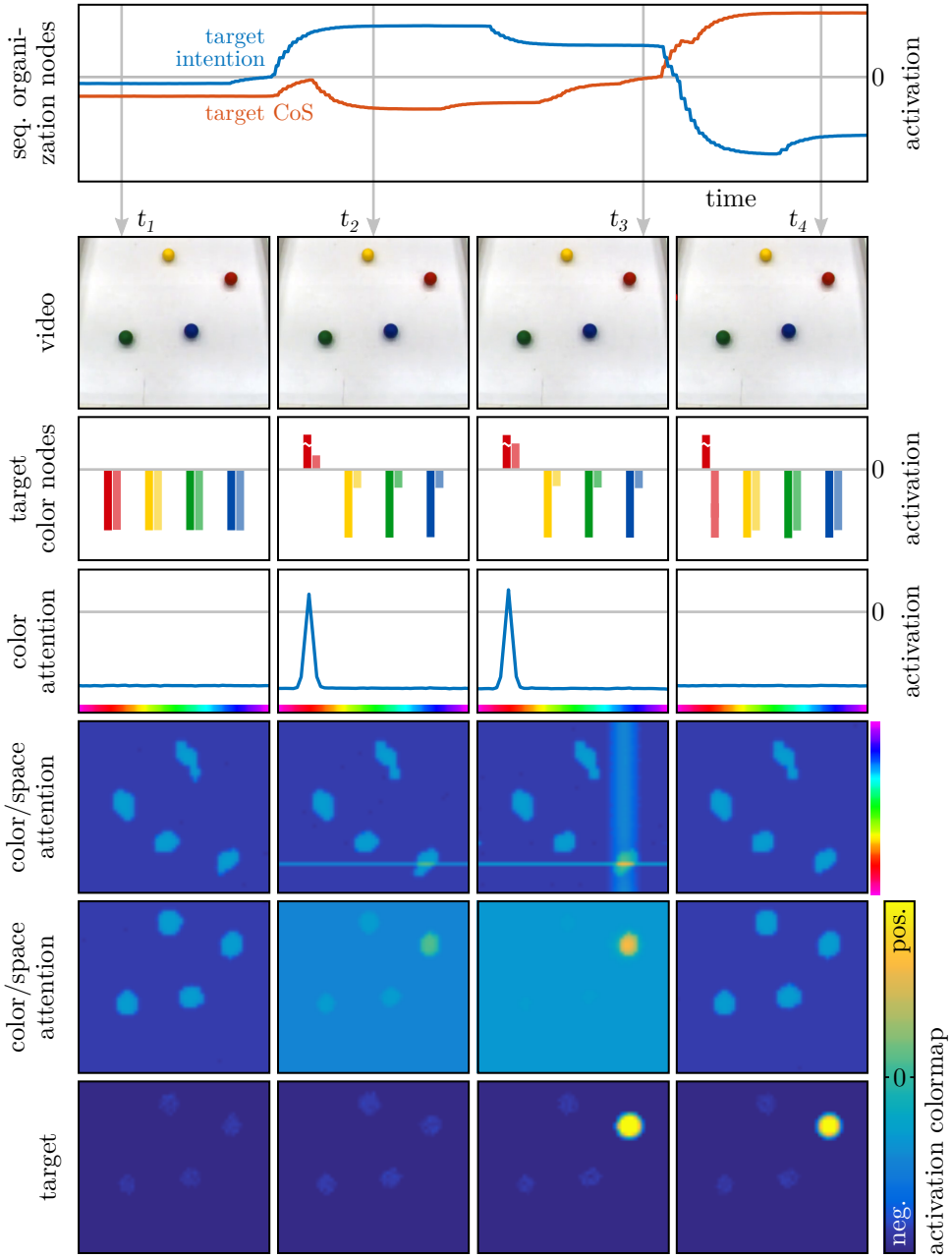

Fig. 7: Grounding the phrase "the red object". The top panel shows the activation levels of the target intention node and its CoS over continuous time. All panels below show the state of the model at four points in time, $t_1, \ldots, t_4$ (four columns). 'Target color' panel: Activation of memory nodes is shown by opaque bars, activation of production nodes by transparent bars. Broken bars denote values that are not to scale. Lower three panels: Activation is color coded (colormap on the right). 'Color/space attention' panel (fifth row): The hue dimension is plotted along the vertical axis. Only the horizontal spatial dimension is shown (along the horizontal axis). 'Color/space attention' panel (sixth row): Only the spatial dimensions are shown. 'Target' panel: Activation in the target field is shown for both spatial dimensions.

This simulation is one out of a set of 18, in which only the color RED of the target object was specified. The visual stimuli varied the number of (red) target objects, the number of (non-red) distractor objects, and the number of moving targets/distractors. In all visual scenes that contained a red object, the model was successful in bringing it into the attentional foreground, thus grounding the phrase. This worked irrespective of the number of distractor objects in the scene and whether they moved or not. Among multiple potential targets in the visual scene, the model selected one, ignoring other candidates. The model prefers red objects that move over stationary red objects because movement enhances salience. Among multiple red objects that are all moving or all stationary, the model typically selects the ones closer to the bottom of the visual scene. This is because the camera axis is slightly inclined so that objects near the bottom are visually larger and thus more salient due to perspective distortion. For all visual scenes that did not contain red objects, the model did not bring any object into the attentional foreground, remaining in its initial state for lack of a mechanism to detect terminal failures of visual search.

A second set of six simulations further assessed guided visual search by specifying only the motion direction, RIGHTWARD, of the target object. The visual stimuli varied the number of target objects (moving rightward) in the scene and the number of distractor objects (stationary or moving somewhere other than rightward). Again, the model finds an object with the matching motion direction whenever there is one, selects one among multiple candidates based on salience, and does not attentionally select an object when there is no matching objects in the visual scene.

4.2 Grounding feature conjunctions

The model can ground objects specified by multiple features, a form of conjunctive feature search. Fig. 8 illustrates an example in which the model grounds the phrase "the red object moving rightward", encoded by activating the respective concept nodes (through CEDAR). The activation snapshots for time t_1 and t_2 are analogous to the previous example of grounding with a single feature. Here, an additional memory node specifies the motion direction RIGHTWARD. At time t_2, the active production nodes for RED and RIGHTWARD (transparent bars, third and fourth row) project (via the color attention field and motion attention field, both not shown) into the color/space attention field and motion/space attention field, respectively. This shows up here as lines of activation along the spatial dimension (row five and six). In the color/space attention field, this line overlaps with input from the two red objects, in the motion/space attention field the line overlaps with the object moving rightward. The projection of these two feature/space fields on the multi-peak spatial attention field and selective spatial attention field (not shown) is strongest at the spatial locations at which these overlaps occur, which can be seen from the activation pattern those fields induce in the target field (bottom row).

The reciprocal interaction between the selective spatial attention field and the two feature/space fields enhances activation at those locations at which both feature/space fields begin to form peaks. This is visible as vertical lines of input in the color/space attention field and motion/space attention field. From this loop of interaction between spatial and feature/space attention, peaks in the feature/space fields emerge at time t_3 at those locations at which matches between the top-down feature cues, red and rightward, coincide. This is how binding through space works in the model. Here, it brings about the attentional selection of the red object moving to the right, visible in the target field at this time: The phrase "the red object moving rightward" has been perceptually grounded.

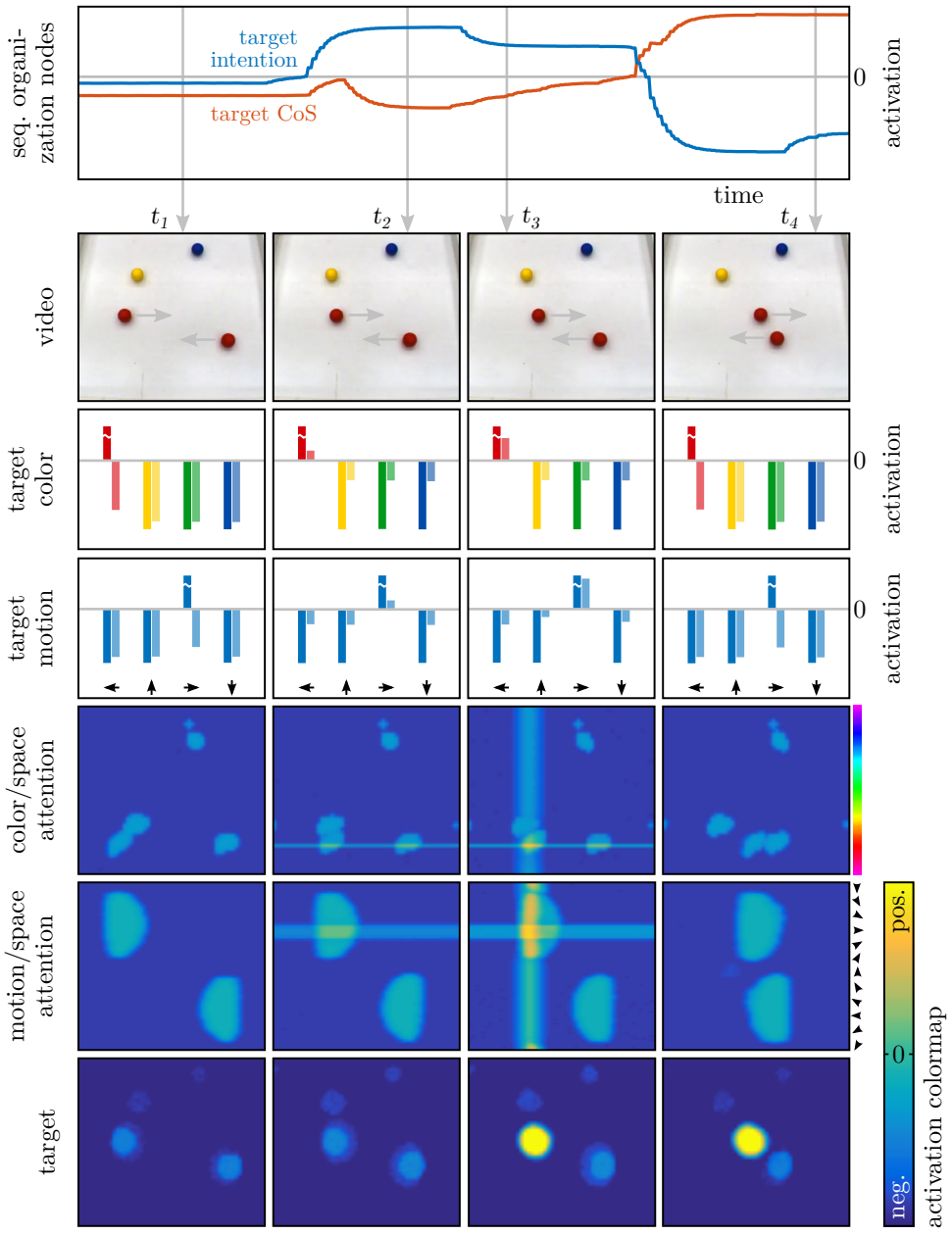

Fig. 8: Grounding the phrase "the red object moving rightward" in a scene with a unique target. The same conventions as in Fig. 7 are used. The plotted nodes and fields differ and are labeled. 'Video' panel: The movement direction of objects is denoted by gray arrows.

Because the CoS of the various intention nodes have been met, activation begins to decay. By time t_4, most of the activation in the model has returned below threshold. What remains active are the target color memory node and the target motion memory node representing the original phrase together with its perceptual grounding in the target field.

This simulation is one out of 33 that probed all qualitatively different cases that may arise in the case of conjunctive feature search. The visual stimuli varied the number of objects in the scene, the number of objects matching the color feature (red), the number of objects matching the feature of motion direction (rightward), and the resulting number of potential target objects. In all cases, the model brought only those objects into the attentional foreground that had the attribute values specified by the feature conjunction. We also systematically checked if objects that match the description only along one feature dimension were selected, for instance, a red object moving leftward, or a green object moving rightward. This is particularly interesting in cases in which multiple objects each match one of the feature values, for instance a scene with a red object moving leftward and a green object moving rightward. In all cases, only objects that match both features were attentionally selected. This demonstrates that the architecture correctly identifies the binding of the features to the same spatial location (Schneegans, Spencer, et al., 2016).

4.3 Grounding relations between objects

We demonstrate how the model grounds relations in an exemplary case that is interesting because it exhibits a form of *hypothesis testing* (Richter et al., 2014). The phrase "the red object to the left of the green object" must be grounded for a scene that contains multiple red and green objects, of which only one pair matches the spatial relation. Analogously to before, Fig. 9 shows activation time courses and snapshots. At time t_1, the phrase has been encoded (through CEDAR) in the target color memory node for RED, the reference color memory node for GREEN, and the spatial relation memory node for TO THE LEFT OF. The grounding process is initiated by giving input to the intention node for finding a target from a relation. The coupling structure of the control system makes that the target (blue line in top panel) and spatial (green line) intention nodes activate first.

By the time t_1, the red object on the right has been brought into the attentional foreground and induced a peak in the target field (fifth row) at its location. Being slightly larger due to the camera geometry, this object has slightly higher saliency. Projection onto the target IOR field (sixth row) induces a self-sustained peak at that location. The spatial relation task node activates the spatial relation production node for the relation TO THE LEFT OF, which projects excitatorily into the spatial relation CoS field (second row from bottom) and inhibitorily into the spatial relation CoD field (bottom row). You can see the spatial pattern of the synaptic connectivity that encodes the concept TO THE LEFT OF in the activation of those fields.

By time t_2, the search for a reference object is active (yellow line in top panel) and has brought the two green objects into the attentional foreground in the reference field (third row from bottom). This enables the coordinate transform that centers activation of the target field on the reference objects (effectively, a superposition of two coordinate transforms) and projects that activation into the spatial relation CoS and CoD fields. Because the candidate red object lies to the right of both green objects, input to the spatial relation CoS field does not overlap with input from the spatial relation production node and no peak is formed (second row from bottom). A peak is, instead, formed in the spatial relation CoD field (bottom row), signalling the rejection of the hypothesis

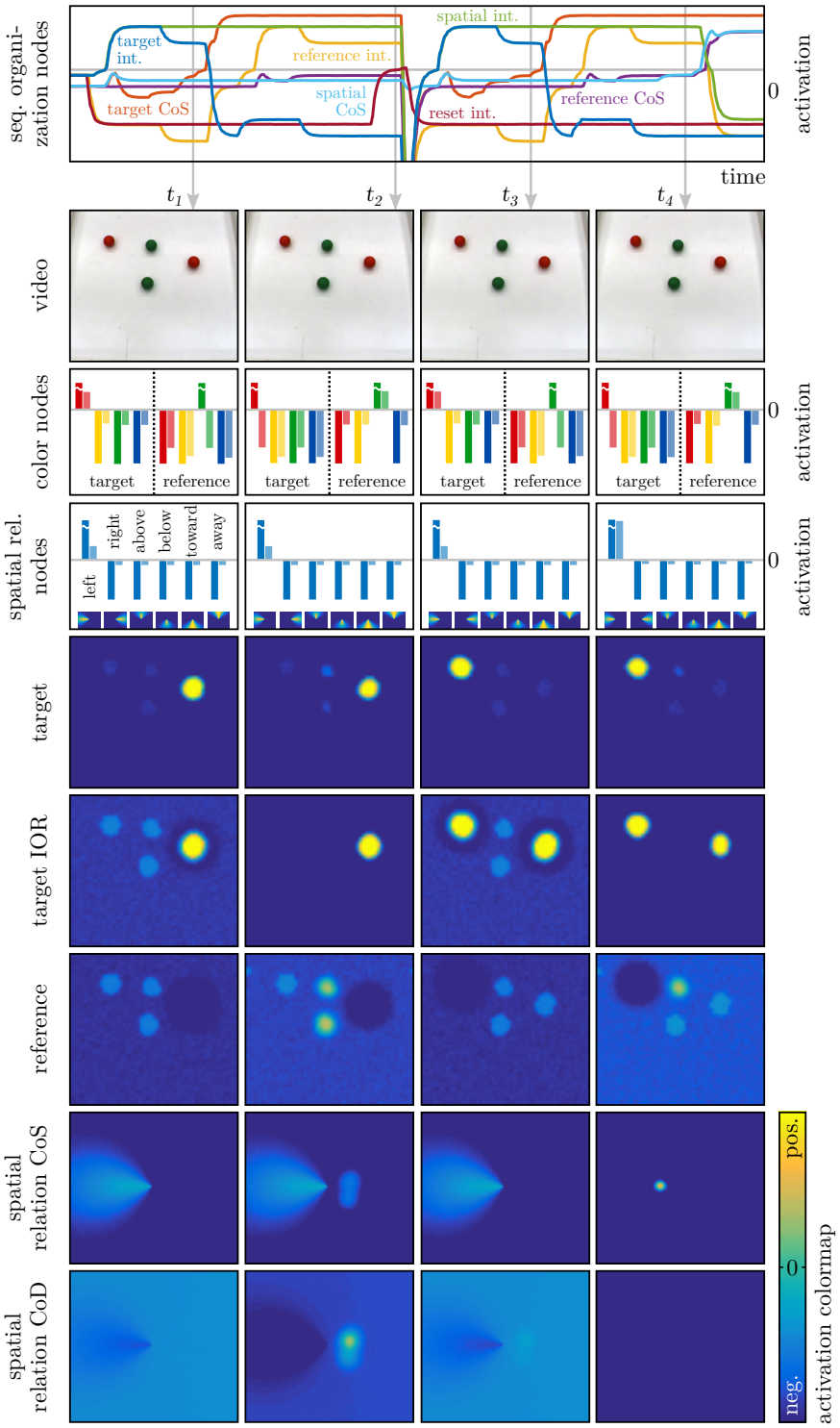

Fig. 9: Grounding the phrase “the red object to the left of the green object” in a scene that requires hypothesis testing. The same conventions as in Fig. 7 are used. The plotted nodes and fields differ and are labeled. ‘Color nodes’ panel: Activation of concept nodes is shown both for the target and reference role.

that the currently selected red object is the correct target. The peak activates the reset intentional node (dark red on top), which, through fixed synaptic connections, inhibits those parts of the model that hold this hypothesis in working memory (top panel after t_2). The grounding process begins anew, starting with target and relation search, but now in the presence of a memory of the first attempt in the target IOR field (fourth row from bottom).

This makes that at time t_3, the leftmost red object has been selected as a candidate target, which is then entered as a second self-sustained peak in the target IOR field. Between t_3 and t_4, the reference search is again activated (yellow line on top) and the two green objects are again brought into the attentional foreground. The coordinate transform projects the current target into the spatial relation CoS and CoD fields, but this time that input matches input from the spatial relation in the spatial relation CoS field.

At time t_4, the spatial relation CoS field has made a selection decision for the upper of the two green objects that better fits the synaptic pattern of the spatial relational concept. The peak in the spatial relation CoS field at that location is transformed back to scene coordinates and projected onto the spatial attention system, where it drives selection of the upper green object, which then shows in the reference field (third row from bottom). At this time, the model has grounded the phrase "the red object to the left of the green object" by generating peaks in the target field, the reference field, and the spatial relation CoS field at the respective locations of the leftmost red object, the upper green object, and the red object's position relative to the green object. All of the model's decisions are based on the fundamental detection and selection instabilities of its component neural fields (see Section 2).

Overall, we performed four sets of simulations with relations between objects. In a first set of eight simulations, task input represented the phrase "the red object to the left of the green object" and there was exactly one red and one green object in the scene. The visual stimuli varied whether or not the spatial configuration of the target and reference object matched the phrase and whether or not the objects were moving. In all eight simulations, the model only brought the pair of objects into the attentional foreground when its spatial relation matched the phrase. The model grounded the phrase irrespective of whether or not the target or reference object were moving (throughout each visual stimulus, the objects' movement did not qualitatively change the spatial relation). When the pair of objects did not match the relation, the model detected the mismatch and rejected the target object, which was held in the inhibition-of-return memory. The model remained in that state indefinitely, waiting for a potential matching target object to appear (as there is no mechanism for "giving up").

In a second set of nine simulations, the same phrase was given, but the target and reference objects were no longer uniquely identifiable by their color. The example described above is from this set. The visual stimuli varied the number of objects in the scene, the number of red and green objects, and, through their spatial configuration, the number of pairs of target and reference that matched the phrase. Again, in all simulations the model only brought pairs of objects into the attentional foreground that matched the specified relation. When there were multiple potential red target or green reference objects, the model selected one in either role so that the pair had the specified relation. The target object is grounded first; its selection among multiple candidates is based on the saliency of the object. The reference object is grounded second; its selection is based both on its saliency and on its match with the specified relation. Pairs of objects that did not match the specified relation were not selected. They were brought into the attentional

foreground transiently, but then suppressed as the model detected the mismatch of their spatial relation.

Two analogous sets of simulations where performed for movement relations, grounding the phrase "the red object moving toward the green object". Again, in both the first set (six simulations), in which there was a single red and green object, and in the second set (nine simulations), in which multiple red or green objects occurred, only pairs matching the relation were grounded. The respective visual stimuli varied analogously to the simulation sets outlined above.

4.4 Describing object attributes

We demonstrate how the model delivers a description of object attributes in a representative example of a visual scene that consists of a single red object moving upward in the camera image (Fig. 10). For this visual scene, we expect the model to activate the memory nodes for RED and UPWARD. Video input is present from the start of the simulation. Shortly after the start of the time line, input given (through CEDAR) to the intention node for describing initiates the process, leading first to the activation of the intention nodes for searching targets (blue line in top panel).

At time t_1, the red object has been selected in the selective spatial attention field, with a peak that tracks the input of the moving object (second row from bottom). Its activation projects onto the color/space attention field and motion/space attention field, visible as vertical lines (fifth and sixth row), and onto the target field.

At time t_2, a peak has formed in the target field, also tracking the moving object, and peaks are forming in the color/space attention field and motion/space attention field.

At time t_3, the color RED and the movement direction UPWARD have been extracted as attributes of the selected object, visible in the activation of the respective production and memory nodes (third and fourth row). By activating these nodes, the model has generated a description of the scene. The active nodes represent the phrase "a red object that is moving upward".

Shortly after time t_3, the CoS node for target search (red line in top panel) is activated, leading to activation throughout the model to decay. At time t_4, what remains active are the peak in the target field, which still tracks the input, and the memory nodes for RED and UPWARD. The intention node for the reference behavior has become active, which enables the description of a reference object and relation if another object were present in the scene. The model does not have a mechanism to detect that there is only a single object in the scene.

In four similar simulations the visual stimuli varied the presence of a single object in the scene, its color, spatial position, and, if moving, its motion direction. Whenever there was an object in the scene, the model correctly extracted all available features of that object (color for stationary objects, color and movement direction for moving objects) and represented them by activating the associated concept node. The model also represented the object's spatial position. When there was no object in the scene, the model did not activate any concept node nor did it activate spatial representations.

4.5 Describing relations between objects

Finally, we illustrate how the model generates a description using a movement relation between two objects in a scene (Fig. 11). The scene consists of two stationary balls, one

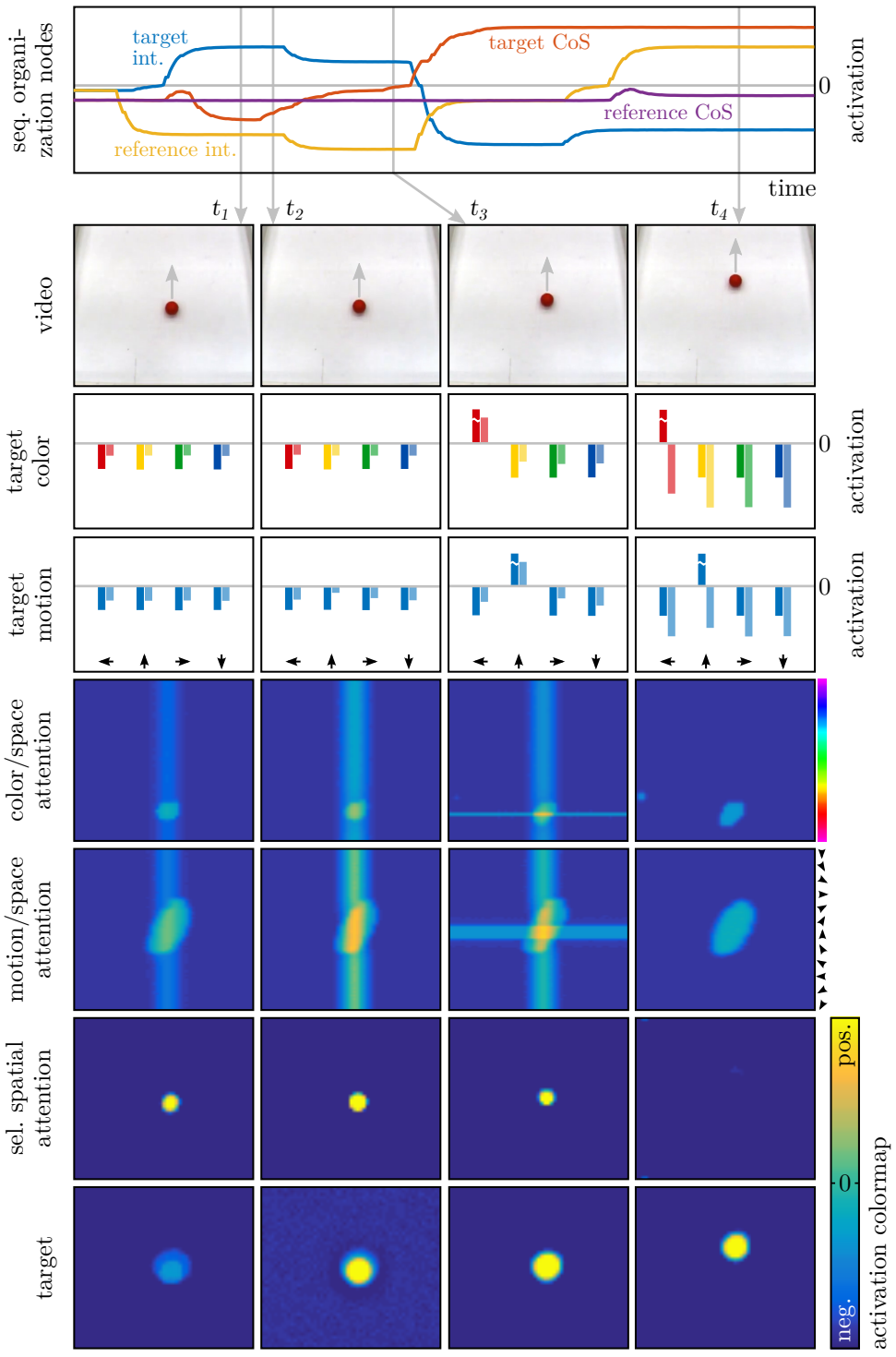

Fig. 10: Describing a scene with a moving red object. The same conventions as in Fig. 7 are used. The plotted nodes and fields differ and are labeled.

blue, one green, and a red ball that is rolling toward the green ball. We expect the model to activate memory nodes representing the phrase "the red object is moving toward the green object". Video input is present from the start of the simulation. Shortly after the start of the time line, input given (through CEDAR) to the intention node for describing initiates the process, leading first to the activation of the intention nodes for searching targets and finding spatial relations (blue and green lines in top panel).

At time t_1, the red object has been selected in the spatial attention field, because its movement makes it the most salience object. This induces a peak in the target field that tracks the position of the red object. Since the model detects movement, the production nodes representing movement relations receive more input than the production nodes representing spatial relations (fourth row from top). The former are activated and project their connectivity patterns onto the spatial relation CoS field (bottom panel).

At time t_2, the color red has been extracted as an attribute of the selected object, visible in the activation of the target color production and memory nodes for RED (third row). The model has also extracted the movement direction of the object and has categorized it as UPWARD (even though the velocity vector is not perfectly vertical) leading to activation of the corresponding target motion production and memory nodes (not shown).

In the meantime, the CoS node for target search (red line in top panel) is activated, leading to deactivation of the target intention node (blue line) and ultimately to activation of the reference intention node (yellow line). By time t_3, all remaining objects in the scene (the blue and green objects) have been brought into the attentional foreground, leading to corresponding peaks in the reference field (third row from bottom). Through the spatial transformations, the relational candidates field (second row from bottom) receives input that reflects the position of the red target object with respect to both reference objects. The projection of the relational candidates field onto the spatial relation CoS field (bottom row) is transformed further by rotation around the center of the field into the direction of the target's motion direction. This leads to the selection of the green reference object, which overlaps better with the input pattern from the two movement relation nodes, specifically, with the lower triangular input from the TOWARD node.

By time t_4, this has led to the selection of the green object as reference in the reference field. Transforming back, the red target's spatial position relative to the green object is stabilized in the relational candidates field. The relational concept TOWARD is activated (fourth row) and the red target object is represented in the spatial relation CoS field relative to the green reference object but rotated into the red object's movement path. The blue object's representations are inhibited.

At the same time, the feature of the selected green reference object has been extracted through the color CoS field based on input from the color/space attention field. This activates the reference color memory node for the color GREEN (third row) via the reference color production node (already back to inactive at time t_4).

Thus, at time t_4, the model has generated a description of the scene by activating the target color memory node for RED, the target motion memory node for UPWARD, the reference color memory node for GREEN, and the spatial relation memory node for the relation TOWARD. That output represents the phrase "a red object that is moving upward and toward a green object".

We performed nine similar simulations, in which the model had to generate a description of a relation between objects. The visual stimuli varied the number of objects in the scene, their spatial configuration, as well as their motion directions relative to each other. In all simulations, whenever the most salient (closest and/or moving) object could be

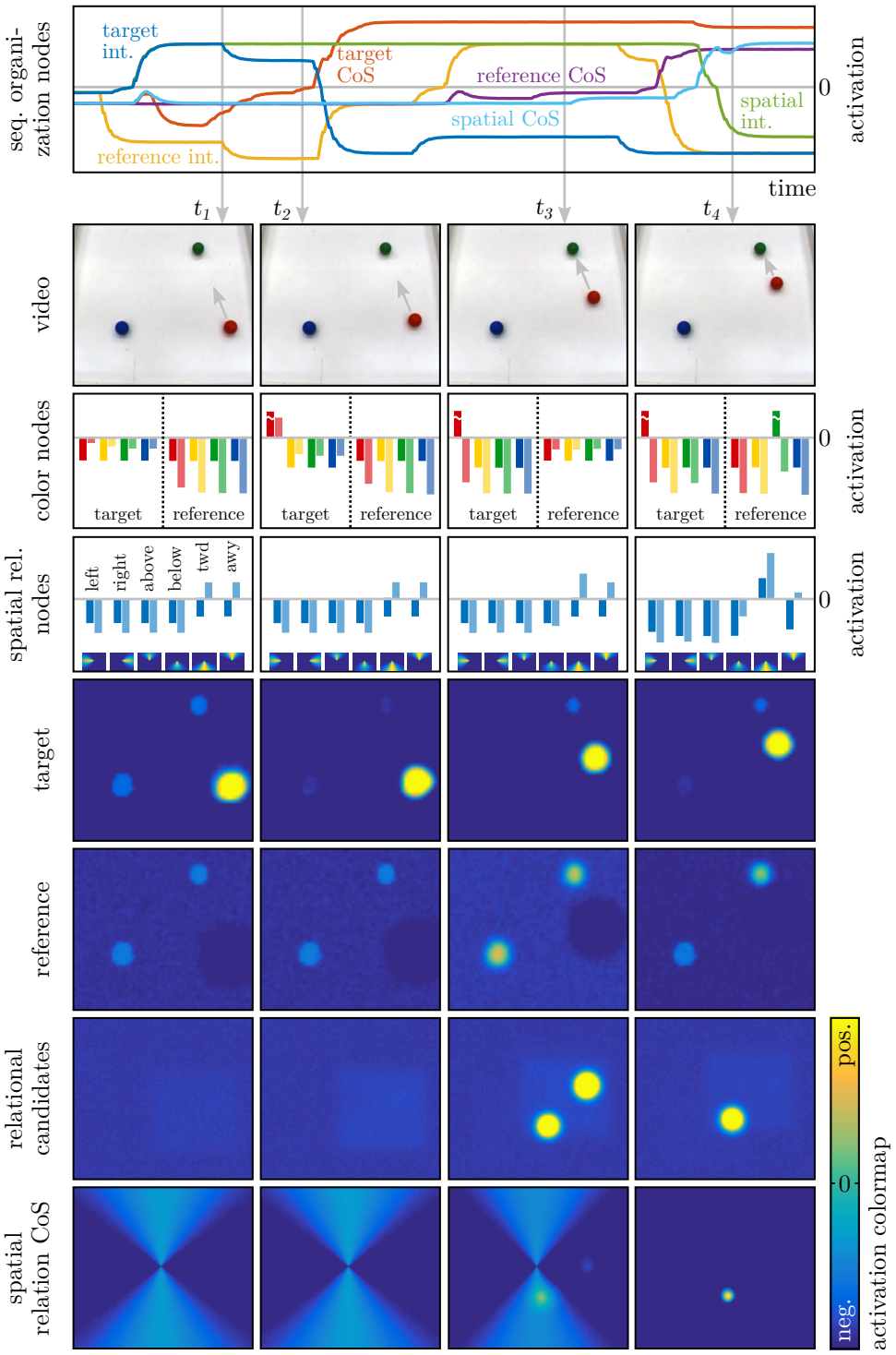

Fig. 11: Activation time courses of the model as it generates a description of the visual event illustrated by snapshots. The same conventions as in Fig. 8 are used. The plotted variables and fields differ and are labeled.

described by a spatial or movement relation, the model correctly selected both a target and a reference object, determined their attributes by activating corresponding feature concept nodes, and identified their relation by activating a relation concept node. When the target object was stationary, the model described it through a spatial relation (e.g., TO THE LEFT OF) with respect to the reference object. When the target object was moving, the model described it using a movement relation (e.g., TOWARD). The model selected the most salient object in the scene as a target. As a reference object, it selected the object that best fit any spatial or movement relation with the target.

When there were multiple ways to describe salient objects in the scene, the model selected a coherent set, so that target object, reference object, and relation matched the scene. The model produced relational phrases whenever there were multiple objects in the scene, even when the most salient target object could have been described unambiguously by its individual color or motion direction attributes alone.

When the model selected a target object and candidates for reference objects, whose relation did not match any of relational concepts implemented in the model, it did not generate a description and selected another object as target instead.

5 Discussion

We have outlined an embodied, neural dynamic account of the perceptual grounding of relations and the autonomous generation of relational descriptions of visual scenes. That account establishes links between phrases and perceptually grounded objects. Phrases are represented by neural activation patterns across concept nodes, while objects are represented by activation patterns across feature maps of visual space. The model builds on a small set of core mechanism of neural processing formalized in DFT. That it does, in fact, deliver the claimed functions is demonstrated in extensive simulations.

5.1 Coordinate transforms for neural function evaluation

Spatial relations may be viewed as functions that take the spatial positions of the target and the reference object as arguments and return a measure of how well these positions match the relation. Substituting other features for the spatial positions, this view captures relations in general. Evaluating functions by referencing their arguments is central to information processing accounts of higher cognition (Anderson, 2007). Understanding how function evaluation can be realized in neural networks is, therefore, a key challenge for neural accounts of higher cognition (Anderson et al., 2008; Kriete et al., 2013; O'Reilly, 2006). In neural networks, functions are instantiated in patterns of connectivity whose inputs deliver the arguments. To neurally implement relations as functions of two arguments, these must be provided in a bound representation that expresses all combinations of values of the two arguments. This was recognized early in connectionist theorizing, most explicitly in the notion of tensor products (Smolensky, 1990) that provide a neuron for every pair that can be formed out of the components of two vectors of neural activity.

Implementing spatial relations in neural networks thus requires a bound representation of all combinations of spatial positions of reference and target objects and connectivity that projects from that bound representation onto a neural representation of relations. This is how recent deep neural networks address the learning of relations (Lu et al., 2016; Santoro et al., 2017). The learned projections extracting relations are copied across visual space by weight sharing. This approach is demanding in the number of connections needed

to instantiate relations, scales badly with the number of relational concepts, and is not neurally plausible.

Our approach shares with this work the notion of a bound representation of the positions of potential reference and target objects. We use that bound representation to implement a single function, however, the active transformation of the visual array into a coordinate frame that is centered on potential reference objects (Lipinski et al., 2012), akin to a simulated gaze shift (Ballard et al., 1997). Each relational concept is then instantiated by a single pattern of connectivity that projects from that transformed representation of visual space to relational concept nodes. Bound representations of gaze and visual space are found in parietal area LIP and elsewhere in the form of gain-field neural populations (for review, see Salinas & Sejnowski, 2001). How gain field neural populations may generate active coordinate transformations is well understood theoretically (Pouget & Sejnowski, 1997; Schneegans & Schöner, 2012). Our approach is thus neurally plausible and provides a neural foundation more broadly for the reference frame transformations recognized as critical to language grounding (Coventry et al., 2018; Franconeri et al., 2012) and to which extant theoretical models refer (Dominey & Boucher, 2005; Gorniak & Roy, 2004; Regier, 1995; Roy, 2005).

In the model, only the neural representation of visual space is actively transformed, while visual features at different spatial locations are referenced in a form of feature binding through space (Schneegans, Lins, et al., 2016; Schneegans, Spencer, et al., 2016). We return to this neural implementation of Feature Integration Theory (Treisman & Gelade, 1980) below. In comparison to the deep network approaches, we note that this fact further reduces the neural resources required to implement relations.

Vector-symbolic architectures (VSA; Plate, 1995; Smolensky, 1990) take a different approach to the binding of variables to enable function evaluation. High-dimensional random neural activation vectors are used as distributed representations of concepts. Because high-dimensional random vectors are almost always orthogonal to each other, they can be combined by operations that do not expand the dimensionality of the representation. This makes it possible to bind and operate on the high-dimensional vectors without losing access to the original component vectors (Levy & Gayler, 2008). In the past, this notion was not credited with neural plausibility as it was unclear how such vectors could be sustained during cognitive processing. The neural-engineering framework (Eliasmith, 2005) has claimed that VSAs could be implemented in neurally plausible spiking neural networks (Eliasmith et al., 2012). The connectivity of these networks must be highly specific, however, to enable the networks to preserve the information originally encoded in the high-dimensional vectors. The neural plausibility of these networks is debatable, therefore. Depending on the outcome of that debate, the VSA approach might either be a radical alternative to the approach outlined here, or it may fail to qualify as a neurally mechanistic approach. It may be too early to settle that issue.

A fundamentally different proposal for how neuronal networks may represent relations builds on the notion of dynamic links (von der Malsburg, 1985). In this conception, neurons encode features or concepts by virtue of their location in a neural network. What neurons represent may then be bound through special mechanisms that link different neurons, such as phase-locked firing patterns (Hummel & Biederman, 1992) or fast synaptic modification that creates links on the fly (see Zylberberg et al., 2013, for review and discussion). A class of neural models (Doumas & Hummel, 2012; Doumas et al., 2008; Martin, 2020) represents relations by dynamically linking neurons representing the relation and its arguments. These models have not, to date, addressed perceptual grounding or the active generation

of relational descriptions, however.

5.2 Autonomous neural processing

Ultimately, the model achieves perceptual grounding of relations by directing visual attention to a target and a reference object in the visual array that match the relation. The component of the model responsible for attentional selection binds feature dimensions through space (Schneegans, Spencer, et al., 2016) in the sense that values along multiple feature dimensions may be combined to attentionally select a relevant object (Grieben et al., 2020). To prevent mis-binding, this can only be done one object at a time, implying the need to sequentially process candidate objects.

The capacity to autonomously generate such sequences of cognitive operations is a strength of the model and new over previous work (e.g., Lipinski et al., 2012). This capacity emerges from the core concepts of DFT, the detection instability and its reverse that accompany the activation and deactivation, respectively, of peaks and nodes. Combined with the sub-network organized around the condition-of-satisfaction (Sandamirskaya & Schöner, 2010), these instabilities generate transitions upon the successful completion of a cognitive operation or, through the condition-of-dissatisfaction, transitions upon the failure to complete a cognitive operation. A hierarchically structured network of neural nodes enables the emergence of complex sequences of processing from the underlying time- and state-continuous neural dynamics.

Rather than invoke algorithms or production rules, the model uses these mechanisms to attend to potential target and reference objects one by one, accepting them as a perceptual grounding of a relation if their spatial configuration matches the relational concept, rejecting them if their spatial configuration does not match the specified relation. An inhibition of return map ensures that the selection of candidate objects explores the visual array.[6]

Although the model as a whole is complex, it is minimal in the sense of covering only the component processes that are necessary to account for perceptual grounding and description generation based on neural principles. This was demonstrated in a set of 104 simulations that systematically probed relational tasks across qualitatively different visual scenes. The capacity of the model to perform across tasks and scenes based on a single set of parameter values shows that there are no inherent conflicts in the model between these different demands that would prevent the model from meeting them all. The parameter values were found by hand, largely constrained by the required dynamic regimes. Our experience suggests that the model's performance is not dependent on these exact parameter values. This illustrates how the model fundamentally differs from curve-fitting type models. Such models are often thought of as a form of data compression, using quantitative criteria to assess their descriptive or predictive power by weighing the number of parameters against the number of data points. This way of assessing models is not well suited to neural process accounts (Schöner et al., 2016, chapter 15).

One interesting alternative is to account for variability from trial to trial, explaining potential errors of grounding or description. Accounts for such variability and errors are possible in DFT models based on the amplification of neural noise near the instabilities that support decision making (Dineva & Schöner, 2018). In the present context, no data of this kind are available, to our knowledge (but see below for work using mouse tracking).

[6]The system may endlessly search for a possible grounding if there there are no objects in the scene that match the description as we have not attempted to address the implied time-out problem.

5.3 Research program

We think of the model as a first step toward an account for relational thinking and higher cognition that takes the constraints of neural processing and embodiment into account, a research program first outlined, perhaps, by Barsalou (1999). We briefly outline the issues, small and large, that such a research program still may have to address.

Feature dimensions Our simulation results are demonstrations that the model can deal with the sampled classes of visual situations. There are many ways how such performance may break down, in particular, as object features such as size, shape, and texture vary. Such failures would primarily reflect properties of the visual front-end, which is kept simple to make the model comprehensible. The visual appearance of the objects themselves was, therefore, kept relatively invariant across the stimulus sets. To deal with more complex scenes, objects, and visual backgrounds, the perceptual component of the model would need to be extended. The underlying neural processes account of conjunctive visual search (Grieben et al., 2020) is scalable, in principle. A larger issue is, however, how neural processes organize visual search when objects are specified by category labels. How neural network models of object recognition that are inspired by the human visual system (Kriegeskorte, 2015; Riesenhuber & Poggio, 1999; Serre et al., 2007) may support visual search is the focus of current research (Zoran et al., 2020).

Reference frames At the core of our theoretical perspective is the assumption that specific reference frames enable the neural instantiation of relations. We have demonstrated how the coordinate transforms for spatial and movement relations may emerge from simple steerable neural maps. These methods can be easily extended to include scale transformations based on the same log-polar maps (Lomp et al., 2017) and transformations to intrinsic reference frames needed, for instance, for relations like BETWEEN (van Hengel et al., 2012). Interesting challenges are reference frames for relations such as NEAR, INSIDE, or TOUCHING, which go beyond the relative spatial position of the center of objects and may involve spatial scales, relations between scales, and representations of boundaries. A classification of relations with respect to the required reference frames may be an important first step.

Learning Understanding how relational concepts may be learned from experience (Doumas et al., 2008) is the goal of an entire research program (Samuelson & Faubel, 2016; Samuelson et al., 2017), to which the neural learning mechanisms of DFT may provide an entry point (Sandamirskaya, 2014; Sandamirskaya & Storck, 2015). In the present model, however, all synaptic connectivity was fixed. The perceptual front-end of the model and the coordinate transformations may be shaped by early development of visual cognition, based on neural substrate that is pre-structured for such functions. The neural machinery for sequentially organizing cognitive operations may likewise be supported by innate substrate and shaped by cognitive development. Given such systems, learning the connectivity pattern for individual concepts and relations from experience may become a well defined theoretical problem (see Tekülve & Schöner, 2020, for an account in a related context). The broader issues of autonomous learning are relevant to all aspects of the architecture, however: How are neural dynamic architectures tuned to ensure that meaningful sequences of cognitive operations emerge? How are the different neural fields tuned to be in an appropriate dynamic regime for a given task? How are instances of

experience detected from which to learn? Clearly, autonomously learning from experience requires neural process infrastructure that organizes learning.

Toward higher cognitive functions The notion of using steered, active coordinate transforms to neurally implement cognitive operators may transfer to cognitive functions that are "higher" in the sense of further removed from perception and action. For instance, the reference task of much work on cognitive architectures (Anderson, 2007), mental arithmetic, may be accessible to neural dynamic thinking based on the neural representation of numerosity (Nieder & Dehaene, 2009). If local patterns of activation represent numbers, coordinate transforming the neural representation of one number steered by the neural representation of another number would enable establishing relations between numbers. Arithmetic operations would be encoded in the connectivity pattern that projects from such transformation fields. Similar ideas may be used to address logical or comparison relations. We are currently using such ideas to study analogical mapping, in which multiple relations between two objects are used to identify a pair with matching relations (Gentner, 2010), all without any explicit symbol manipulation.

Sequences of relations Scaling the approach toward relational thinking and language would require to deal with sequences of relations. The model grounds or generates one relational phrase at a time. Even that simplest case already entails the problem of transferring the outcome of a first operation, a grounding attempt of a target object, to a second operation, the grounding attempt of the reference object. That transfer occurs through the pattern of activation in the perceptual representations. Grounding or generating sequences of phrases requires more generally that outcomes are transferred across phrases, a form of argument passing that is not easy to achieve in neural networks. In preliminary work we have shown how the patterns of activation in the grounded perceptual representations may serve this function if they are sustained in the form of a mental map (Sabinasz et al., 2020). Such an extension must also address the neural machinery necessary for controlling sequences of cognitive operations beyond the minimal two-step sequences we have demonstrated.

Embodiment The model uses real video streams as input, which illustrates that it fulfils the embodiment constraint at the perceptual end. How would the model drive actual motor behavior? DFT models have routinely addressed motor tasks (Knips et al., 2017; Tekülve et al., 2019), which requires many more component processes, of course. Such extensions may be useful if motor behavior could provide evidence for the postulated neural processes of perceptual grounding. We have made a first step toward such evidence (Lins & Schöner, 2019). In a variant of the mouse tracking paradigm (Spivey & Dale, 2006), participants moved the mouse-controlled cursor onto the target designated by a relational phrase. During movement, the mouse trajectory was attracted to the location of the reference object. Given the close link between visual attention and the spatial requirements of motor acts (Baldauf & Deubel, 2010), this finding is consistent with the model bringing both the reference and the target object into the attentional foreground. Related work in the visual world paradigm is broadly consistent with our postulate of sequential attentional selection of target and reference objects (Burigo & Knoeferle, 2015).

Toward communication Perceptual grounding and generating descriptions are minimal elements of human communication about shared environments (Tenbrink et al., 2017).

Linking phrases to perception in both directions is a strength of our approach, as typical models cover only either production or comprehension (Pickering & Garrod, 2013). A key limitation of our account is, however, the salience driven attentional selection of objects for description. In real communicative settings, speakers take into account many other dimensions such as shared experience and the listener's perspective (Talmy, 2017), or the combination of gesture with language (MacNeill, 2000). The DFT framework provides entry points for such dimensions by supporting biased competition, but a neural process account for many of these factors is a major research challenge. In preliminary work, we have demonstrated that the neural grounding machinery presented here may support the building of neural maps (Kounatidou et al., 2018), perhaps a first step toward an understanding of how shared representations may emerge.

6 Conclusion

The model we described here demonstrates in a concrete setting how the processes of perceptual grounding and the generation of descriptions may be explained based on the fundamental principles of neural dynamics. More generally, it demonstrates how elements of higher cognition, including the capacity to sequentially operate on objects, to form and reject hypotheses, and to apply local networks to global representations, may be grounded in sensory-motor processing.

References

Amari, S.-i. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, *27*(2), 77–87. https://doi.org/10.1007/bf00337259

Anderson, J. R. (2007). *How can the human mind occur in the physical universe*. New York, Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195324259.001.0001

Anderson, J. R., Fincham, J. M., Qin, Y., & Stocco, A. (2008). A central circuit of the mind. *Trends in Cognitive Sciences*, *12*, 136–143. https://doi.org/10.1016/j.tics.2008.01.006

Baldauf, D., & Deubel, H. (2010). Attentional landscapes in reaching and grasping. *Vision Research*, *50*(11), 999–1013. https://doi.org/10.1016/j.visres.2010.02.008

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*(4), 723–767. https://doi.org/10.1017/s0140525x97001611

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–609, discussion 610–60. https://doi.org/10.1017/S0140525X99002149

Berger, M., Faubel, C., Norman, J., Hock, H., & Schöner, G. (2012). The counter-change model of motion perception: An account based on dynamic field theory (A. E. P. Villa, Ed.). In A. E. P. Villa (Ed.), *ICANN 2012, Part I, LNCS 7552*, Berlin Heidelberg, Springer. https://doi.org/10.1007/978-3-642-33269-2_73

Burigo, M., & Knoeferle, P. (2015). Visual attention during spatial language comprehension. *PLoS ONE*, *10*(1), 1–21. https://doi.org/10.1371/journal.pone.0115758

Clark, A. (1999). An embodied cognitive science? *Trends in Cognitive Sciences*, *3*(9), 345–351. https://doi.org/10.1016/s1364-6613(99)01361-3

Cohen, M. R., & Newsome, W. T. (2009). Estimates of the contribution of single neurons to perception depend on timescale and noise correlation. *Journal of Neuroscience*, *29*(20), 6635–6648. https://doi.org/10.1523/jneurosci.5179-08.2009

Coventry, K. R., Andonova, E., Tenbrink, T., Gudde, H. B., & Engelhardt, P. E. (2018). Cued by what we see and hear: Spatial reference frame use in language. *Frontiers in Psychology*, *9*, 1–14. https://doi.org/10.3389/fpsyg.2018.01287

Dineva, E., & Schöner, G. (2018). How infants' reaches reveal principles of sensorimotor decision making. *Connection Science*, *30*(1), 53–80. https://doi.org/10.1080/09540091.2017.1405382

Dominey, P. F., & Boucher, J. D. (2005). Learning to talk about events from narrated video in a construction grammar framework. *Artificial Intelligence*, *167*(1-2), 31–61. https://doi.org/10.1016/j.artint.2005.06.007

Douglas, R. J., & Martin, K. A. C. (2004). Neural circuits of the neocortex. *Annual Review of Neuroscience*, *27*, 419–451. https://doi.org/10.1146/annurev.neuro.27.070203.144152

Doumas, L. A. A., & Hummel, J. E. (2012). Computational models of higher cognition (K. J. Holyoak & R. G. Morrison, Eds.). In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199734689.013.0005

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*(1), 1–43. https://doi.org/10.1037/0033-295X.115.1.1

Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nature Neuroscience*, *3*(November), 1184–1191. https://doi.org/10.1038/81460

Eliasmith, C. (2005). A unified approach to building and controlling spiking attractor networks. *Neural Computation*, *17*, 1276–1314. https://doi.org/10.1162/0899766053630332

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, C., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, *338*(6111), 1202–1205. https://doi.org/10.1126/science.1225266

Erlhagen, W., Bastian, A., Jancke, D., Riehle, A., & Schöner, G. (1999). The distribution of neuronal population activation (DPA) as a tool to study interaction and integration in cortical representations, *94*(1), 53–66. https://doi.org/10.1016/s0165-0270(99)00125-9

Faugeras, O., Touboul, J., & Cessac, B. (2009). A constructive mean-field analysis of multi-population neural networks with random synaptic weights and stochastic inputs. *Frontiers in Computational Neuroscience*, *3*(February), 1–28. https://doi.org/10.3389/neuro.10.001.2009

Franconeri, S. L., Scimeca, J. M., Roth, J. C., Helseth, S. A., & Kahn, L. E. (2012). Flexible visual processing of spatial relationships. *Cognition*, *122*(2), 210–227. https://doi.org/10.1016/j.cognition.2011.11.002

Gärdenfors, P. (2014). *The geometry of meaning: Semantics based on conceptual spaces*. The MIT Press. https://doi.org/10.1017/S0332586515000268

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, *34*(5), 752–775. https://doi.org/10.1111/j.1551-6709.2010.01114.x

Gibbs, R. W., & Colston, H. L. (1995). The cognitive psychological reality of image schemas and their transformations. *Cognitive Linguistics*, *6*(4), 347–378. https://doi.org/10.1515/cogl.1995.6.4.347

Gorniak, P., & Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, *21*, 429–470. https://doi.org/10.1613/jair.1327

Grieben, R., Tekülve, J., Zibner, S. K., Lins, J., Schneegans, S., & Schöner, G. (2020). Scene memory and spatial inhibition in visual search—A neural dynamic process model and new experimental evidence. *Attention, Perception, & Psychophysics*, *82*, 775–798. https://doi.org/10.3758/s13414-019-01898-y

Henson, R. N. A., & Burgess, N. (1997). Representations of serial order (J. A. Bullinaria, D. W. Glasspool, & G. Houghton, Eds.). In J. A. Bullinaria, D. W. Glasspool, & G. Houghton (Eds.), *4th Neural Computation and Psychology Workshop, London 9-11 April 1997: Connectionist Representations*, London, UK, Springer.

Hummel, J. E., & Biederman, I. (1992). Dynamic Binding in a Neural Network for Shape Recognition. *Psychological Review*, *99*(3), 480–517. https://doi.org/10.1037/0033-295X.99.3.480

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203. https://doi.org/10.1038/35058500

Jackendoff, R. (2012). *A user's guide to thought and meaning*. Oxford, UK, Oxford University Press.

Jancke, D., Erlhagen, W., Dinse, H. R., Akhavan, A. C., Giese, M., Steinhage, A., & Schöner, G. (1999). Parametric population representation of retinal location: Neuronal interaction dynamics in cat primary visual cortex. *Journal of Neuroscience*, *19*(20), 9016–9028. https://doi.org/10.1523/JNEUROSCI.19-20-09016.1999

Johnson, J. S., Spencer, J. P., Luck, S. J., & Schöner, G. (2009). A dynamic neural field model of visual working memory and change detection. *Psychological Science*, *20*(5), 568–77. https://doi.org/10.1111/j.1467-9280.2009.02329.x

Knips, G., Zibner, S. K. U., Reimann, H., Popova, I., & Schöner, G. (2017). A neural dynamics architecture for grasping that integrates perception and movement generation and enables on-line updating. *Frontiers in Neurorobotics*, *11*(9). https://doi.org/10.1109/IROS.2014.6942627

Kounatidou, P., Richter, M., & Schöner, G. (2018). A neural dynamic architecture that autonomously builds mental models (T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish, Eds.). In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, Cognitive Science Society.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Kriete, T., Noelle, D. C., Cohen, J. D., & O'Reilly, R. C. (2013). Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences*, *110*(41), arXiv arXiv:1408.1149, 16390–16395. https://doi.org/10.1073/pnas.1303547110

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. New York, Basic Books.

Landau, B., & Jackendoff, R. (1993). What and where in spatial language and spatial cognition. *Behavioral and Brain Sciences*, *16*, 217–265. https://doi.org/10.1017/S0140525X00029733

Langacker, R. W. (1986). An introduction to cognitive grammar. *Cognitive Science*, *10*, 1–40. https://doi.org/10.1207/s15516709cog1001_1

Levy, S. D., & Gayler, R. (2008). Vector symbolic architectures: A new building material for artificial general intelligence, In *Proceedings of the 2008 Conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, Amsterdam, The Netherlands, IOS Press. http://dl.acm.org/citation.cfm?id=1566174.1566215

Lins, J., & Schöner, G. (2019). Computer mouse tracking reveals motor signatures in a cognitive task of spatial language grounding. *Attention, Perception, & Psychophysics*, *81*(7), 2424–2460. https://doi.org/10.3758/s13414-019-01847-9

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neurobehavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1490–1511. https://doi.org/10.1037/a0022643

Logan, G. D., & Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations (P. Bloom, M. Peterson, L. Nadel, & M. Garrett, Eds.). In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and Space*. Cambridge, MA, USA, MIT Press.

Lomp, O., Faubel, C., & Schöner, G. (2017). A neural-dynamic architecture for concurrent estimation of object pose and identity. *Frontiers in Neurorobotics*, *11*(April), 1–17. https://doi.org/10.3389/fnbot.2017.00023

Lomp, O., Richter, M., Zibner, S. K. U., & Schöner, G. (2016). Developing dynamic field theory architectures for embodied cognitive systems with cedar. *Frontiers in Neurorobotics*, *10*, 1–18. https://doi.org/10.3389/fnbot.2016.00014

Lu, C., Krishna, R., Bernstein, M., & Fei-Fei, L. (2016). Visual relationship detection with language priors, In *European Conference on Computer Vision*, Springer International Publishing. https://doi.org/10.1007/978-3-319-46448-0_51

MacNeill, D. (Ed.). (2000). *Language and gesture*. Cambridge University Press. https://doi.org/10.1017/CBO9780511620850

Martin, A. E. (2020). A compositional neural architecture for language. *Journal of Cognitive Neuroscience*, 1–20. https://doi.org/10.1162/jocn_a_01552
doi: 10.1162/jocn_a_01552

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*(8), 348–56. https://doi.org/10.1016/j.tics.2010.06.002

Nieder, A., & Dehaene, S. (2009). Representation of number in the brain. *Annual Review of Neuroscience*, *32*(1), 185–208. https://doi.org/10.1146/annurev.neuro.051508.135550

O'Reilly, R. C. (2006). Models of high-level cognition. *Science*, *314*, 91–94. https://doi.org/10.1126/science.1127242

Panzeri, S., Macke, J. H., Gross, J., & Kayser, C. (2015). Neural population coding: Combining insights from microscopic and mass signals. *Trends in Cognitive Sciences*, *19*(3), 162–172. https://doi.org/10.1016/j.tics.2015.01.002

Pastra, K., & Aloimonos, Y. (2012). The minimalist grammar of action. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1585), 103–117. https://doi.org/10.1098/rstb.2011.0123

Perko, L. (2001). *Differential Equations and Dynamical Systems* (3rd). Berlin, Springer Verlag.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347. https://doi.org/10.1017/S0140525X12001495

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, *6*(3), 623–641. https://doi.org/10.1109/72.377968

Pouget, A., & Sejnowski, T. J. (1997). Spatial transformations in the parietal cortex using basis functions. *Journal of Cognitive Neuroscience*, *9*(2), 222–237. https://doi.org/10.1162/jocn.1997.9.2.222

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, *6*(July), 576–582. https://doi.org/10.1038/nrn1706

Regier, T. (1995). A model of the human capacity for categorizing spatial relations. *Cognitive Linguistics*, *6*(1), 63–88. https://doi.org/10.1515/cogl.1995.6.1.63

Richter, M. (2018). *A neural dynamic model for the perceptual grounding of spatial and movement relations* (PhD thesis). Ruhr-Universität Bochum. https://nbn-resolving.org/urn:nbn:de:hbz:294-60740

Richter, M., Lins, J., Schneegans, S., Sandamirskaya, Y., & Schöner, G. (2014). Autonomous neural dynamics to test hypotheses in a model of spatial language (P. Bello, M. Guarini, M. McShane, & B. Scassellati, Eds.). In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, Austin, TX, Cognitive Science Society.

Richter, M., Sandamirskaya, Y., & Schöner, G. (2012). A robotic architecture for action selection and behavioral organization inspired by human cognition, In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, New York, NY, Institute of Electrical; Electronics Engineers (IEEE). https://doi.org/10.1109/iros.2012.6386153

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–25. https://doi.org/10.1038/14819

Roy, D. (2005). Grounding words in perception and action: Computational insights. *Trends in Cognitive Sciences*, *9*(8), 389–396. https://doi.org/10.1016/j.tics.2005.06.013

Roy, D. (2008). A mechanistic model of three facets of meaning (M. de Vega, A. Glenberg, & A. Graesser, Eds.). In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and Embodiment: Debates on Meaning and Cognition*. Oxford, UK, Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199217274.003.0011

Rutishauser, U., Douglas, R. J., & Slotine, J.-J. (2010). Collective stability of networks of winner-take-all circuits. *Neural Computation*, *773*1105.3106, 735–773. https://doi.org/10.1162/NECO_a_00091

Sabinasz, D., Richter, M., Lins, J., & Schöner, G. (2020). Grounding spatial language in perception by combining concepts in a neural dynamic architecture (S. Denison, M. Mack, Y. Xu, & B. C. Armstrong, Eds.). In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42th Annual Conference of the Cognitive Science Society*, Cognitive Science Society.

Salinas, E., & Sejnowski, T. J. (2001). Gain modulation in the central nervous system: Where behavior, neurophysiology, and computation meet. *The Neuroscientist*, *7*(5), 430–440. https://doi.org/10.1177/107385840100700512

Samuelson, L. K., & Faubel, C. (2016). Grounding word learning in space and time (G. Schöner, J. Spencer, & the DFT Research Group, Eds.). In G. Schöner, J. Spencer, & the DFT Research Group (Eds.), *Dynamic thinking: A primer on dynamic field theory*. New York, Oxford University Press.

Samuelson, L. K., Kucker, S. C., & Spencer, J. P. (2017). Moving word learning to a novel space: A dynamic systems view of referent selection and retention. *Cognitive Science*, *41*, 52–72. https://doi.org/10.1111/cogs.12369

Sandamirskaya, Y. (2014). Dynamic neural fields as a step toward cognitive neuromorphic architectures. *Frontiers in Neuroscience*, *7*(276), 1–13. https://doi.org/10.3389/fnins.2013.00276

Sandamirskaya, Y., & Schöner, G. (2010). An embodied account of serial order: How instabilities drive sequence generation. *Neural Networks*, *23*(10), 1164–1179. https://doi.org/10.1016/j.neunet.2010.07.012

Sandamirskaya, Y., & Storck, T. (2015). Learning to look and looking to remember: A neural-dynamic embodied model for generation of saccadic gaze shifts and memory formation (P. Koprinkova-Hristova, V. Mladenov, & N. K. Kasabov, Eds.). In P. Koprinkova-Hristova, V. Mladenov, & N. K. Kasabov (Eds.), *Artificial Neural Networks*. Springer International Publishing. https://doi.org/10.1007/978-3-319-09903-3_9

Santoro, A., Raposo, D., Barrett, D. G. T., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017). A simple neural network module for relational reasoning. *arXiv e-prints*, arXiv 1706.01427.

Schneegans, S., Lins, J., & Spencer, J. P. (2016). Integration and selection in multidimensional neural fields (G. Schöner & J. P. Spencer, Eds.). In G. Schöner & J. P. Spencer (Eds.), *Dynamic Thinking: A Primer on Dynamic Field Theory*. New York, Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199300563.003.0005

Schneegans, S., & Schöner, G. (2012). A neural mechanism for coordinate transformation predicts pre-saccadic remapping. *Biological Cybernetics*, *106*(2), 89–109. https://doi.org/10.1007/s00422-012-0484-8

Schneegans, S., Spencer, J. P., & Schöner, G. (2016). Integrating "what" and "where": Visual working memory for objects in a scene (G. Schöner & J. P. Spencer, Eds.). In G. Schöner & J. P. Spencer (Eds.), *Dynamic Thinking: A Primer on Dynamic Field Theory*. New York, Oxford University Press.

Schöner, G. (2019). The dynamics of neural populations capture the laws of the mind. *Topics in Cognitive Science*, 1–15. https://doi.org/10.1111/tops.12453

Schöner, G., Spencer, J. P., & the DFT Research Group. (2016). *Dynamic thinking: A primer on dynamic field theory*. New York, Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199300563.001.0001

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–457. https://doi.org/10.1017/S0140525X00005756

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), 411–426. https://doi.org/10.1109/TPAMI.2007.56

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1-2), 159–216. https://doi.org/10.1016/0004-3702(90)90007-M

Spencer, J. P., & Schöner, G. (2003). Bridging the representational gap in the dynamic systems approach to development. *Developmental Science*, *6*(4), 392–412. https://doi.org/10.1111/1467-7687.00295

Spivey, M. J., & Dale, R. (2006). Continuous dynamics in real-time cognition. *Current Directions in Psychological Science*, *15*(5), 207–211. https://doi.org/10.1111/j.1467-8721.2006.00437.x

Talmy, L. (1988). The relation of grammar to cognition (B. Rudzka-Ostyn, Ed.). In B. Rudzka-Ostyn (Ed.), *Topics in Cognitive Linguistics*. Amsterdam/Philadelphia, John Benjamins. https://doi.org/10.1075/cilt.50.08tal

Talmy, L. (2017). *The targeting system of language*. Cambridge MA, USA, MIT Press.

Tekülve, J., Fois, A., Sandamirskaya, Y., & Schöner, G. (2019). Autonomous sequence generation for a neural dynamic robot: Scene perception, serial order, and object-oriented movement. *Frontiers in Neurorobotics*, *13*, 95. https://doi.org/10.3389/fnbot.2019.00095

Tekülve, J., & Schöner, G. (2020). A neural dynamic network drives an intentional agent that autonomously learns beliefs in continuous time. *IEEE Transactions on Cognitive and Developmental Systems*, 1–12.

Tenbrink, T., Andonova, E., Schole, G., & Coventry, K. R. (2017). Communicative success in spatial dialogue: The impact of functional features and dialogue strategies. *Language and Speech*, *60*(2), 318–329. https://doi.org/10.1177/0023830916651097

Tomasello, M. (1995). Joint attention as social cognition (J. C. Moore & P. J. Dunham, Eds.). In J. C. Moore & P. J. Dunham (Eds.), *Joint Attention: Its Origins and Role in Development*, Lawrence Erlbaum Associates, Inc.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

van Hengel, U., Sandamirskaya, Y., Schneegans, S., & Schöner, G. (2012). A neural-dynamic architecture for flexible spatial language: Intrinsic frames, the term "between", and autonomy, In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, IEEE. https://doi.org/10.1109/ROMAN.2012.6343746

von der Malsburg, C. (1985). Nervous Structures With Dynamical Links. *Berichte der Bunsengesellschaft/Physical Chemistry Chemical Physics*, *89*(6), 703–710.

Wilson, H. R., & Cowan, J. D. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Biological Cybernetics*, *13*(2), 55–80. https://doi.org/10.1007/BF00288786

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–36. https://doi.org/10.3758/bf03196322

Wu, W., Tiesinga, P. H., Tucker, T. R., Mitroff, S. R., & Fitzpatrick, D. (2011). Dynamics of population response to changes of motion direction in primary visual cortex. *Journal of Neuroscience*, *31*(36), 12767–77. https://doi.org/10.1523/JNEUROSCI.4307-10.2011

Zoran, D., Chrzanowski, M., Huang, P.-S., Gowal, S., Mott, A., & Kohl, P. (2020). Towards robust image classification using sequential attention models, In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE. http://arxiv.org/abs/1912.02184

Zylberberg, A. D., Paz, L., Roelfsema, P. R., Dehaene, S., & Sigman, M. (2013). A neuronal device for the control of multi-step computations. *Papers in Physics*, *5*(July), arXiv 1312.6660, 1–14. https://doi.org/10.4279/PIP.050006