An integrated cognitive model of visuospatial relation tracking

Sangeet Khemlani[1], Branden Bio[2], and J.G. Trafton[1]
sangeet.s.khemlani.civ@us.navy.mil, brandenbio@gmail.com, greg.j.trafton.civ@us.navy.mil
[1]U.S. Naval Research Laboratory, Washington, DC 20375 USA
[2]University of Wisconsin-Madison, Madison, WI 53705 USA

Abstract

We developed a computational cognitive theory of visuospatial relation tracking to model how people perceive and monitor spatial relations such as *left of* and *next to*. Previous work suggests that they build mental simulations of their environment, or of imagined or described scenarios, to reason about space and spatial concepts – but no account explains how visual information is used to construct and update those simulations. Our model converts perceptual data (objects and locations detected by a convolutional neural network) into an integrated sparse iconic simulation of the scene – a perceptual model. Perceptual models are cognitively plausible representations that are more noise tolerant and stable than raw percepts. They allow for efficient latent encoding of high-level relations. We tested the cognitive model against a benchmark dataset for spatial relation recognition, and show a close fit between human and model-based perception of 2D spatial relations.

Keywords: spatial relation tracking; spatial perception; perceptual models; spatial models; mReasoner; simulation

Introduction

Newborns can perceive spatial relations: after habituating to a blinking square on the right side of a display, they shift their attention to a blinking square on the left (Gava, Valenza, & Turati, 2009). And adults routinely integrate their perception of objects relative to one another into their high-level thinking and reasoning. Imagine you see a blue car parked to the left of a green car. Somebody comments: "The blue car is next to the green car." You understand immediately that the description is accurate by integrating your observation with what you know about the meaning of the phrase "next to." Many cognitive scientists argue that humans coordinate linguistic and perceptual sensemaking by constructing mental simulations of objects in the environment in a manner akin to how they are perceived (Burge, 2022; Dantzig & Pecher, 2011; Carey, 2009; Hegarty, 2004; Knauff, 2013; Johnson-Laird, 2006a; Tversky, 2005; but cf. Quilty-Dunn, 2016). Some have argued further that human inferential processes operate over these mental simulations – and *not* on the imagistic information derived from perception, and likewise *not* on propositional content derived from language (Johnson-Laird, Byrne, & Khemlani, 2024). Yet no integrated cognitive model of the process exists.

Cognitive models of spatial processing are targeted in scope: they seek to explain the underlying mental processes humans carry out when interpreting or perceiving spatial relations. For example, Ragni and Knauff's (2013) PRISM model of spatial reasoning explains the piecemeal operations of interpreting and processing spatial relations. It argues that interpretation processes yield predictable preferences in the structures of the spatial simulations that reasoners build, which explains common reasoning errors, illusory inferences, and the relative difficulty people exhibit when drawing conclusions about different spatial reasoning problems. The theory doesn't explain how people perceive relations from visual input – it isn't designed to do so. But other investigators, such as Franconeri and colleagues, explore the computations that underlie the process of perceiving relations. Their "attentional shift" account argues that people perceive whether one object is to the left of another by shifting attentional selection in the corresponding direction (Franconeri et al., 2012; cf. Regier & Carlson, 2001). The theory explains why attentional shifts can facilitate relational memory (Yuan, Uttal, & Franconeri, 2016), and how spatial relations are encoded and processed asymmetrically (Roth & Franconeri, 2012). But it isn't designed to explain how people reason deductively or inductively from spatial information – and so it cannot explain reasoning biases or difficulties. Both PRISM and attentional shift theory exemplify two aspects of models of spatial cognition, and perhaps cognitive modeling in general: first, they tend to implement theories with narrow ambits that focus on specific human behaviors, e.g., locating objects in imagery or drawing conclusions from verbal premises. They are not general-purpose accounts of spatial cognition. Second, their purpose is to provide new insights into cognitive processes, and their value rests in their ability to predict and explain empirical phenomena, as well as their theoretical conciseness and breadth (Cassimatis, Bello, & Langley, 2008).

In contrast, contemporary AI systems – such as vision-language models (VLMs; see e.g., Li et al., 2015; Zhang, Huang, Jin, & Lu, 2024) – can jointly process imagistic and linguistic input. Many such systems are "end-to-end" architectures in which a single model is trained to handle every part of a processing pipeline, from perception to deduction. VLMs decompose images into patches, then interpret those patches as tokens represented as vectors embedded in a high-dimensional space; and they convert text to tokens embedded in a different high-dimensional space. Tasks are represented textually, in another high-dimensional space, and bridging layers are used to coalesce and draw correspondences between visual, linguistic, and task embeddings, while attentional layers learn associations of how images relate to words. The resulting systems can be remarkably general – they can interpret and process spatial natural language queries such as, "find the object that's bigger than a toaster," which is a kind of query that has received little empirical investigation by experimental

psychologists (cf. Majid et al., 2004). Yet they are not suitable for serving as models of human cognition for at least three reasons. First, end-to-end architectures make it difficult to derive testable theoretical predictions from intermediate stages of processing. They make few proposals for how the mind processes perceptual input or tracks intermediate representations. Second, their visual and linguistic systems have no shared, common representational format akin to a mental simulation of an environment – and so their bridging layers are not guaranteed to yield the same systematicity that human mental representations provide. And as result, they fail at many rudimentary tasks (Kamath, Hessel, & Chang, 2023; Liu, Emerson, & Collier, 2023; Yu et al., 2025). For example, many VLMs have difficulty recognizing spatial converses; are unable to draw hypothetical inferences from imagery, and are disrupted by irrelevant surface features of prompts (Khemlani et al., 2025; Tran, Khemlani, & Trafton, 2025). Third, even in limited cases where empirical predictions can be gleaned from their operations, those predictions can be brittle too, and they do not reflect the constraints of any capacity-limited cognitive architecture (see Bowers et al., 2025; Orr et al., 2025 for consonant arguments).

This paper accordingly introduces a cognitive theory that serves as an integrated high-level account of perception and interpretation processes. It describes the theory's computational implementation in the mReasoner cognitive model of reasoning, and shows how it provides a strong fit to the data from a high-powered conceptual replication of a benchmark experiment on spatial perception. It concludes by highlighting the theoretical and computational advances of the model.

An integrated theory of spatial perception, interpretation, and reasoning

How do people track the spatial relations in the scenes they observe? One broad framework of thinking and reasoning argues that the seemingly unrelated cognitive processes of visual perception, language interpretation, and imagination all produce the same underlying mental representation: a discrete model of the situation (Hegarty, 2004; Johnson-Laird, 2006b; Knauff, 2013; Tversky, 2005). The account argues that reasoners build spatial models when observing or understanding arrangements of objects and scenes and they scan and manipulate those models to make inferences about possible, necessary, or likely outcomes (Byrne & Johnson-Laird, 1989). The theory rests on several core assumptions:

- **Spatial models bear an iconic structure: they mirror the situations they represent.** Spatial models use mental tokens for individual objects and entities in perceived or imagined situations. They arrange those tokens in a way that reflects their visual or physical arrangement. Spatial models are therefore iconic (Peirce, 1931-1958, Vol. 4) as far as possible. They integrate abstract, non-visualizable information – such as negation (see Khemlani et al., 2012) – using symbolic tags on iconic structures (Bio & Khemlani, 2026). Reasoners can manipulate spatial models through piecemeal transformations to make inferences about dynamic sequences (Johnson-Laird et al., 2022; Khemlani et al., 2013).
- **Models are coherent**, i.e., they cannot contain self-contradictory elements. For example, a spatial model can represent three tokens – *A*, *B*, and *C* – in any arrangement, but their structures prohibit reasoners from inferring both *A is to the left of B* and *A is not to the left of B* from a single model. To imagine hypothetical and counterfactual scenarios, i.e., those that run counter to observation and factual knowledge, reasoners have to construct separate internally coherent models (Byrne, 2016; Johnson-Laird, Byrne, & Khemlani, 2024; Khemlani et al., 2018).
- **Inferences emerge from a model's structure.** Spatial relations and deductions are consequences that emerge from the iconic structures of models; reasoners draw conclusions by scanning and revising models (Goodwin & Johnson-Laird, 2005), and not by recourse to some form of symbolic logical calculus (see Johnson-Laird et al., 2024; cf. Rips, 1994). Inferences are easy when they emerge from the structures of models of perception, or those built by default; they are more difficult when reasoners have to revise models and consider multiple alternatives.

Studies on how people interpret spatial relations described in verbal premises, such as "the blue car is next to the green car" (Carrieras & Santamaria, 1997; Byrne & Johnson-Laird, 1989) corroborate the model theory's predictions on the kinds of difficulties reasoners face when reasoning about space (Jahn et al., 2007; Ragni et al., 2016). No existing theory explains how people coordinate their inductive and deductive inferences with their higher-level perceptions, but the model theory's underlying algorithms can, in principle, apply to both linguistic and perceptual input (Johnson-Laird & Khemlani, 2023). A viable theory of visuospatial relation processing should, at a minimum, explain how people track spatial relations in observations. To explore the model theory's capacity for spatial relation tracking, we extended a computational implementation of the model theory with the capacity to build spatial models from real-world imagery. Below we explain how the theory serves as a plausible account for spatial relation perception and tracking.

Spatial relation tracking in mReasoner

mReasoner is cognitive model that serves as a unified computational implementation of the principles of the model theory (Khemlani & Johnson-Laird, 2022; Johnson-Laird & Khemlani, 2013, 2023). It operates by relying on two separate reasoning pipelines: one that builds a single initial model and draw an inference from that model alone, and another that considers alternative models and counterexamples to interrogate its initial inference. Hence, the system implements a dual-process theory of reasoning (Evans & Stanovich, 2013; Kahneman, 2011). It explains deductive inference across a variety of domains, including inferences about quantities (Khemlani & Johnson-Laird, 2022), causality (Briggs & Khemlani, 2019; Khemlani et al., 2018b),

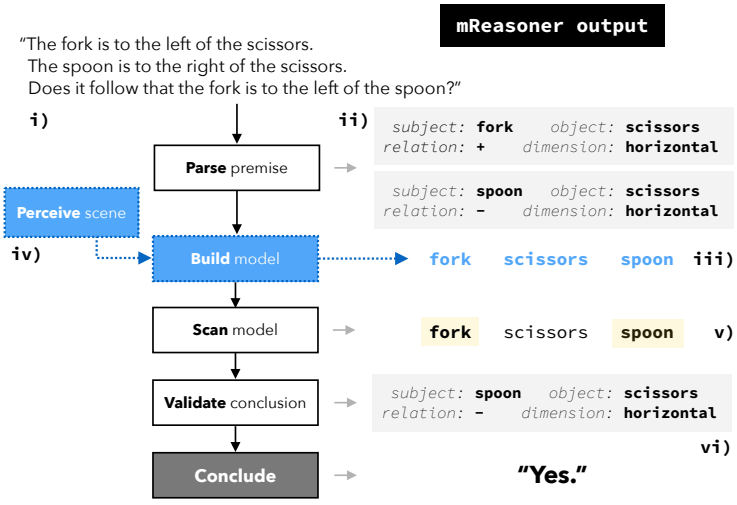

Figure 1. Multimodal spatial model construction in mReasoner. The system can parse natural language assertions (i) to extract their intensional semantics (ii). It can build a spatial model (iii) from these intensions – or build one directly from perception (iv; see elements in blue). It can scan the model (v) to locate tokens in relation to one another, and validate that a relation holds between those tokens (vi).

compound sentences (Khemlani et al., 2018a), and spatiotemporal relations (Kelly, Khemlani & Johnson-Laird, 2020; Khemlani, 2022). To date, mReasoner operated by parsing and building models from premises in a subset of natural language. For instance, it can interpret these premises:

> The fork is to the left of the scissors.
> The spoon is to the right of the scissors.

to build a model as seen in Figure 1. The figure provides a schematic summary of how the system builds models (i-iii), scans them (v), and validates conclusions (vi). A novel component of the system (iv) can construct spatial models directly from perception. We explain its algorithm below.

Building perceptual models from imagery

We developed a perceptual layer that constructs spatial models from images depicting 2D spatial layouts (see Figure 2). The algorithm relies on convolutional neural network – YOLO v10 (Redmon et al., 2016; Sapkota et al., 2015) – to produce a set of object labels and their bounding-box dimensions. The system operates by applying a dynamic grid over each image or frame of video. For each cell in the grid, the system searches through bounding boxes to compute whether the majority of the object's area falls within the cell. If it does, then the system places the corresponding object label in the corresponding cell. Multiple objects can be placed in a cell at once. After each detected object is placed in an appropriate cell, the matrix of objects is simplified to drop irrelevant rows and columns. The resulting representation serves as a coherent, iconic spatial model of the environment, i.e., a perceptual model of space. This algorithm for transducing imagery is designed to convert cognitively implausible representations (digital images and video) into cognitively plausible representations, i.e., perceptual models. And it serves as a proxy for a more robust, attention-oriented cognitive architecture (Kotseruba & Tsotsos, 2025).

Once the system constructs a perceptual model of a frame of video, the model is sent to mReasoner, which scans each token in the model to check whether the model satisfies the semantic constraints of one or more spatial relations. For instance, mReasoner can parse the model in Figure 2 to detect any of the following relations:

> The fork is to the left of the spoon.
> The scissors are to the left of the spoon.
> The scissors are in between the spoon and the fork.
> The leftmost object is the fork.

as well as their converses (e.g., *the rightmost object is the spoon*).

Relation tracking parameters

mReasoner's perceptual layer (Figure 2) includes three parameters that govern how it encodes raw imagery into perceptual models, as follows:

1. **`cellsize`** controls the size of the cells of a grid overlayed onto a image; smaller cellsizes are less likely to group two objects in the same cell.
2. **`compression`** controls the probability of skipping over empty cells in a 2D layout. In the example in Figure 2, if there is a large space between the fork and the scissors, a model can track that empty space (`compression = 0.0`)

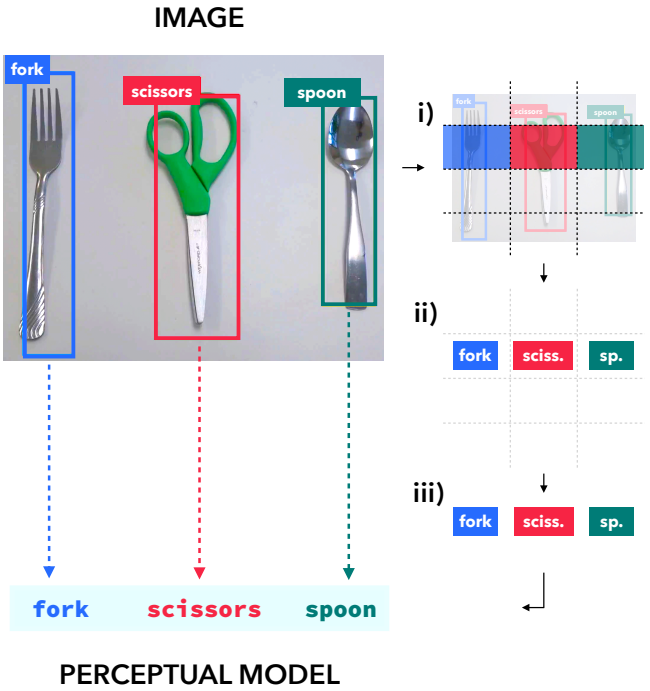

Figure 2. mReasoner's perception layer uses YOLOv10 to find bounding boxes of objects in an image; it then overlays a parameterized grid over an image, and computes the cell in which the central mass of an image falls (i); an object label is then placed in the corresponding cell as a tracker (ii); the underlying image is no longer necessary, and empty cells in the grid are discarded to create a perceptual model (iii).

or else compress the model so that no space is tracked (`compression = 1.0`).

3. **`distance`** controls the way mReasoner scans and validates relations, such that `distance = 1` means that mReasoner will only validate relations when they are adjacent to a referent object, whereas `distance = ∞` validates relations no matter how far they are from a referent object.

The **`cellsize`** parameter affects the way images are transduced; it serves as a proxy for how individual objects can be agglomerated and treated as part of an ensemble (Alvarez, 2011), and it provides the system with the functionality of extracting fine-grained relational distinctions in highly crowded images. The other two parameters affect how models are constructed and scanned. They permit variation in model construction and scanning competence, and they are designed to simulate how reasoners construct preferred perceptual models (cf. Jahn et al., 2007).

To evaluate the system's ability to track spatial relations, we collected data on a study that serves as a high-powered conceptual replication of a classic study on how people describe and interpret spatial imagery in terms of spatial language (Hayward & Tarr, 1995). We describe that study below, and describe how mReasoner fits its data in the section that follows.

Benchmarking experiment

Hayward and Tarr (1995) ran a series of studies on how reasoners track spatial relations in 2D imagery. They presented a reference object in the center of the screen (see Figure 3, left) and a target object in some position other than the center. In one study, participants produced descriptions of the image (Experiment 1); in another, they rated how applicable one of four prototypical spatial relations (*above*, *below*, *left*, and *right*) was to an image (Experiment 2). Their studies were remarkable in their simplicity, and showed that for a given spatial predicate such as *left*, there existed a prototypical region in space relative to a reference object for which the predicate was applicable.

Despite these foundational results, the raw data from Hayward and Tarr's (1995) study are unavailable; the study was underpowered (20 participants in Experiment 2); and the experimental task placed ecologically questionable constraints on participants (i.e., it is unlikely that observers should assess a given object the center of a visual display). We therefore ran a high-powered conceptual replication of their Experiment 2. The study similarly assessed the ability to track relations in 2D imagery, but the images provided to participants were more abstract (circles identified by letters; see Figure 3, right). And participants' task was more straightforward: they merely had to answer the question, "Is the X ball [above/below/to the left of/to the right of] the Y ball?" They responded "yes" or "no" by pressing buttons provided on the screen.

Results replicated the overarching patterns of acceptability highlighted in Hayward and Tarr (1995) and served as a benchmark dataset for relational tracking.

Method

Participants. A total of 102 participants were recruited through Amazon Mechanical Turk. Two participants failed to answer attention checks and were dropped, yielding 100 participants that were subjected to analysis (45 females, 55 males, mean age = 45.02).

Task, design, and materials. Participants received images depicting two circles ("balls") labeled by letters, arranged at seemingly random positions on the screen (see Figure 3). Each ball was 35 px in diameter. They also received a prompt describing the two balls, such as, "Is the H ball above the W ball?" and they responded by clicking buttons marked "Yes" or "No". The experiment varied the spatial relation in the prompt, which, following Hayward and Tarr (1995), asked whether a given ball was "above", "below", "to the left of", or "to the right of" another. Ball locations were assigned randomly, but with the following constraint. A given ball's location was either close to (within 70-140 px), near (within 140-210 px) or far (within 210-280 px) from the other. Participants carried out each prompt for two randomly assigned locations given these constraints, and so they carried out a total of 24 trials, which varied the relation (left, right, above, below) × the distance (near, middle, far) × two instances of the problem.

Procedure. Participants were familiarized with the experiment and given a practice trial; each carried out the 24 trials in a randomized order. The positions of the response buttons were randomized on each trial, and names of balls were assigned randomly. Each participant received two attention check questions, which appeared just like any other trial but probed their understanding of the task (i.e., "Are there 2 balls on the screen?" and "Is there a ball with the number 2 on it?"). The questions had definitive correct answers ("yes" and "no", respectively).

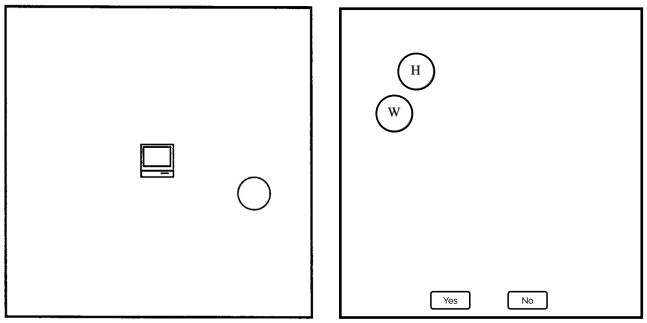

Figure 3. *Left panel:* example display used in Hayward and Tarr (1995), in which a computer icon was presented in the center of a screen and a circle presented at some other point. *Right panel:* materials used in the benchmark experiment; participants saw two circles labeled with letters, and were asked about a particular spatial relation (e.g., "Is the H ball above the W ball?").

	Spatial relation			
Distance	*left*	*right*	*above*	*below*
near	94	90	94	92
middle	90	94	94	95
Far	96	95	98	94
	93	*93*	*95*	*94*

Table 1. Classification accuracies (%s) as a function of the four spatial relations in the benchmarking experiment and the distances between objects on each trial; marginal means are provided (in *italics*) for the four relations.

Open science. Data, materials, experimental code, mReasoner code, and synthetic data derived from computational simulations are available at **[link omitted]**.

Results and validation

To assess data quality, data were coded for accuracy relative to mathematical formalisms that embody their semantics (see Freeman, 1974). They reveal that participants were highly systematic and accurate (relative to those formalisms) at classifying relations: they produced responses that matched mathematical formalisms 94% of the time; these patterns held for the separate conditions of the experiment (see Table 1 for specific relation × distance breakdowns). And they replicate Hayward and Tarr's (1995) overall relation acceptability ratings. In sum, these data serve as a viable benchmark for classifying spatial relations. We omit further analyses of the data for brevity, in favor of describing the mReasoner's simulations of the two studies.

Simulation

The benchmarking experiment yielded a total of 2400 (24 problems × 100 participants) unique problems. To simulate them, we ran a grid search by varying three relevant parameters in mReasoner's perception layer (Busemeyer & Diederich, 2010) The parameter settings were quantized to span their ranges as follows:

```
   cellsize: 70, 85, 100, 115, 130
   distance: 2, 4, 6, 8, 10
compression: 0.0, 0.2, 0.4, 0.6, 0.8, 1.0
```

These variations yielded 150 separate parameter settings. For each cell in the grid search, we generated 100 simulated participants assuming that each participant took all 24 unique problems in the benchmark study, which yielded 360,000 data points. This dataset was then fit against the results of the benchmarking study to locate optimal parameter values, which were as follows, namely where `cellsize = 70`, `distance = 4`, and `compression = 0.0`. These parameter values suggest minimal cellsize, moderate distance, and no compression whatsoever. We used them to generate 1,000 synthetic datapoints for each of the 2400 problems in the benchmarking study. We observed a close fit between data and simulation, both overall (r = .94, RMSE = .25), and separately for each of the different relations in the dataset (*left:* r = .96, RMSE = .18; *right:* r = .94, RMSE = .17; *above:* r = .93, RMSE = .17; *below:* r = .95, RMSE = .16). Figure 4 presents a comparison of the classification of *left* for the 2400 problems in the study. Results reveal a close fit between the data (in blue) and mReasoner's simulations (in red).

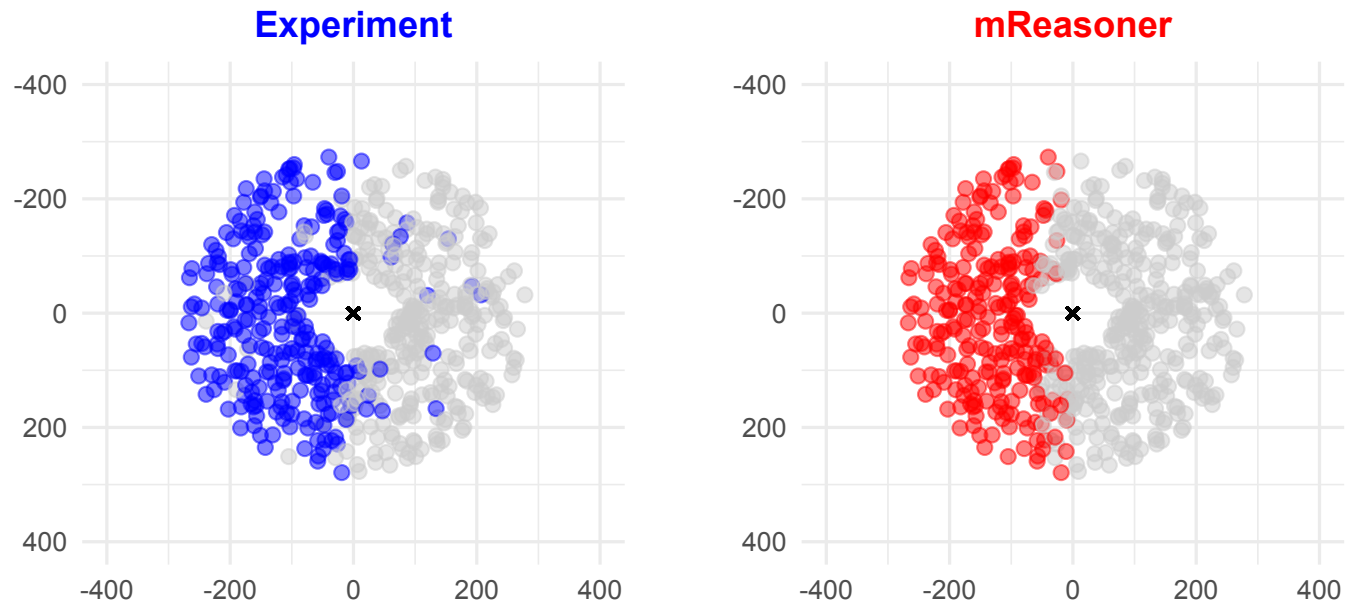

Figure 4. Responses to the question, "Is the Y-ball is to the left of the X-ball?" for 2400 different problem variations, as produced by participants in the benchmarking experiment (*left panel*) and by simulated participants generated by mReasoner (*right panel*). For each problem, the randomized location of the referent object (e.g., the X-ball) was treated as the center (with axes that denote pixels in Cartesian space), and the different points around that object denote various locations of the target object (e.g., the Y-ball). Gray points are those for which participants answered "no". Blue points denote human "yes" responses and red points denote simulated "yes" responses. Results replicate Hayward and Tarr's (1995) thesis that spatial relations refer to prototypical spatial regions.

General discussion

Theorists have argued that reasoners track spatial relations in the scenes they observe by constructing small-scale mental simulations of their environment (Burge, 2022; Carey, 2009; Johnson-Laird, 2006). They build similar such simulations when comprehending spatial language. This project implemented a computational proof-of-concept that such a system was viable, more cognitively plausible, and more robust than existing AI architectures. It developed a new perceptual layer of an existing computational model of human reasoning, mReasoner, which until now could only process verbal input (Khemlani & Johnson-Laird, 2022). The perceptual layer interacts with existing model-building mechanisms to convert images into iconic models of perception. It does so by adapting a convolutional neutral network for object detection to locate objects and their dimensions; it then overlays a parameterized grid and places each detected object in a single cell on that grid, and discards empty rows and columns to yield an iconically structured representation. The underlying representation abstracts over irrelevant imagistic information and simplifies computations for assessing relations; it therefore accords with recent proposals that reasoners are non-committal in their mental imagery (Bigelow, McCoy, & Ullman, 2023).

We tested the computational model by running a high-powered conceptual replication of a classic study on spatial perception and language by Hayward and Tarr (1995), which probed participants' understanding of spatial relations such as "left" and "right". The model carried out the same problems provided to participants across parameterized conditions that simulated variations in how scenes are scanned and perceptual models are constructed. It yielded a close fit to the benchmark data. The resulting model provides a system capable of heterogeneous reasoning (Johnson-Laird, 2006a), i.e., reasoning over both linguistic and perceptual input. For instance, in the introductory example above, a reasoner can verify whether the sentence "the blue car is next to the green car" is true or false by constructing a spatial model of its observations and scanning that model – and not raw imagery – to efficiently assess whether the two are adjacent to one another: perceptual models are more noise tolerant and stable than raw percepts, because minor variations in the positions of the vehicles can yield the same perceptual model.

Some critics argue that the outputs of perception are non-iconic (e.g., Quilty-Dunn, 2016; see also Clarke, 2019). Quilty-Dunn (2016) notes that labels on objects help the mind track the same object over different percepts, and that these labels are syntactic and not iconic in structure. The critique is borne out by the modeling framework we describe above, which reduces a percept into a label, and its therefore consonant with Quilty-Dunn's pluralist view that perceptual output can carry both iconic and non-iconic information. Nevertheless, as we show, the labels themselves can be structured in an iconic manner, and these iconic structures permit rapid inference and perceptual tracking.

Because the algorithm we devised converts imagery into iconic spatial models, it yields certain emergent patterns of relation tracking that may map to how humans update their representations of visual scenes online. Indeed, procedures for maintaining and updating simulations of environments often require specialized calculi, but the present architecture carries out many such procedures using no more resources than those demanded by its transduction and model-processing algorithms. The system can update perceptual representations in three separate ways. First, it can perform *additive updating*, i.e., it can keep track of when a new object enters or leaves a scene. Because perceptual models carry relational information efficiently, the system can automatically track the new relations that the new object bears to the other objects already in the scene. It can also perform *relational updating*, in which the relations between different objects in a scene shift over time. It predicts and explains relational discontinuities in scene perception (Logan & Sadler, 1996; Lovett & Franconeri, 2017), i.e., "jumps" in human interpretations of scenes based on small shifts in how objects are arranged. Finally, the system can perform *agglomerative updating*: it can group items to infer that they're "in the same place as" one another, even though typical object detection networks and perceptual systems rarely produce identical bounding boxes for separate objects. That behavior is an emergent property of the way the system permits multiple objects to occur in the same location.

In sum, we describe an integrated model of visuospatial relation tracking. It serves as a cognitively plausible framework for how perception interacts with higher-order cognitive processes, such as inductive and deductive inference.

Acknowledgements

This work was supported a grant from the Naval Research Lab, and data were collected by Knexus Research Corporation. The research benefitted from feedback and critiques from Bill Adams, Neha Bhatt, David Beltrán, Gordon Briggs, Ruth Byrne, Tony Harrison, Phil Johnson-Laird, Markus Knauff, Maria Kon, Andrew Lovett, Gabe Parnell, and Marco Ragni.

References

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, 15.

Bigelow, E. J., McCoy, J. P., & Ullman, T. D. (2023). Non-commitment in mental imagery. *Cognition*, 238, 105498.

Bio, B. J., & Khemlani, S. (2026). Naïve epistemics: A theory of rational and error-prone mental state reasoning. *Cognition*, 271, 106328.

Bowers, J. S., Puebla, G., Thorat, S., Tsetsos, K., & Ludwig, C. J. (2025). Centaur: A model without a theory. *OSF*.

Briggs, G. & Khemlani, S. (2019). A cognitively plausible algorithm for casual inference. In T. Stewart (Ed.), *Proceedings of the 17th International Conference on Cognitive Modeling*.

Burge, T. (2022). *Perception: First form of mind.* Oxford University Press.

Byrne, R. M. (2016). Counterfactual thought. *Annual Review of Psychology*, 67, 135-157.

Byrne, R. M., & Johnson-Laird, P. N. (1989). Spatial reasoning. *Journal of Memory and Language*, 28, 564-575.

Carey, S. (2009). *The origin of concepts.* New York: OUP.

Carreiras, M., & Santamaría, C. (1997). Reasoning about relations: Spatial and nonspatial problems. *Thinking & Reasoning*, 3.

Cassimatis, N. L., Bello, P., & Langley, P. (2008). Ability, breadth, and parsimony in computational models of higher-order cognition. *Cognitive Science*, 32, 1304-1322.

Dantzig, S. V., & Pecher, D. (2011). Spatial attention is driven by mental simulations. *Frontiers in Psychology*, 2, 40.

Evans, J.S., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*, 223-241.

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.

Freeman, J. (1975). The modelling of spatial relations. *Computer Graphics and Image Processing*, 4, 156-171.

Gava, L., Valenza, E., & Turati, C. (2009). Newborns' perception of left–right spatial relations. *Child Development*, *80*, 1797-1810.

Goodwin, G. P., & Johnson-Laird, P. N. (2005). Reasoning about relations. *Psychological Review*, 112, 468.

Hayward, W. G., & Tarr, M. J. (1995). Spatial language and spatial representation. *Cognition*, 55, 39-84.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*, 280-285.

Jahn, G., Knauff, M., & Johnson-Laird, P. N. (2007). Preferred mental models in reasoning about spatial relations. *Memory & Cognition*, 35.

Johnson-Laird, P. N. (2006a). Models and heterogeneous reasoning. *Journal of Experimental & Theoretical Artificial Intelligence*, *18*, 121-148.

Johnson-Laird, P.N. (2006b). *How we reason.* NY: OUP.

Johnson-Laird, P. N., Bucciarelli, M., Mackiewicz, R., & Khemlani, S. (2022). Recursion in programs, thought, and language. *Psychonomic Bulletin & Review*, *29*.

Johnson-Laird, P.N., Byrne, R.M.J., & Khemlani, S. (2024). Models of possibilities instead of logic as the basis of human reasoning. *Minds and Machines*, 34, 19.

Johnson-Laird, P.N., & Khemlani, S. (2023). Mental models and the algorithms of deduction. In R. Sun (Ed.), *Cambridge Handbook of Computational Cognitive Sciences*.

Kahneman, D. (2011). *Thinking, fast and slow.* Macmillan.

Kamath, A., Hessel, J., & Chang, K. W. (2023). What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 9161-9175).

Kelly, L., Khemlani, S., & Johnson-Laird, P.N. (2020). Reasoning about durations. *Journal of Cognitive Neuroscience*, 32, 2103-2116.

Khemlani, S. (2022). A computational cognitive theory of temporal reasoning. In V. Vandekerckhove, T. Stewart, and D. Kellen (Eds.), *International Conference on Cognitive Modeling 2022*. Society for Mathematical Psychology.

Khemlani, S., Byrne, R.M.J., & Johnson-Laird, P.N. (2018a). Facts and possibilities: A model-based theory of sentential reasoning. *Cognitive Science, 42*, 1887-1924.

Khemlani, S., & Johnson-Laird, P.N. (2013). The processes of inference. *Argument & Computation*, 4, 1-20.

Khemlani, S., & Johnson-Laird, P.N. (2022). Reasoning about properties: A computational theory. *Psychological Review.*

Khemlani, S., Mackiewicz, R., Bucciarelli, M., & Johnson-Laird, P.N. (2013). Kinematic mental simulations in abduction and deduction. *Proceedings of the National Academy of Sciences*, 110.

Khemlani, S., Tran, T., Gyory, N., Harrison, A. M., Lawson, W. E., Thielstrom, R., ... & Trafton, J. G. (2025). Vision language models are unreliable at trivial spatial cognition. arXiv preprint arXiv:2504.16061.

Khemlani, S., Orenes, I., & Johnson-Laird, P.N. (2012). Negation: a theory of its meaning, representation, and use. *Journal of Cognitive Psychology*, 24.

Khemlani, S., Wasylyshyn, C., Briggs, G., & Bello, P. (2018b). Mental models and omissive causation. *Memory & Cognition*, 46.

Knauff, M. (2013). *Space to reason: A spatial theory of human thought*. MIT Press.

Kotseruba, I., & Tsotsos, J. K. (2025). *The computational evolution of cognitive architectures.* Oxford University Press.

Logan, G. D., & Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. A. Peterson, L. Nadel, & M. F. Garrett (Eds.), *Language and space* (pp. 493–529). The MIT Press.

Li, Z., Wu, X., Du, H., Nghiem, H., & Shi, G. (2025). Benchmark evaluations, applications, and challenges of large vision language models: A survey. arXiv preprint arXiv:2501.02189.

Liu, F., Emerson, G., & Collier, N. (2023). Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11.

Lovett, A., & Franconeri, S. L. (2017). Topological relations between objects are categorically coded. *Psychological Science*, 28.

Majid, A., Bowerman, M., Kita, S., Haun, D. B., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, 8, 108-114.

Orr, M., Cranford, D., Ford, K., Gluck, K., Hancock, W., Lebiere, C., ... & Stocco, A. (2025). Not even wrong: On the limits of prediction as explanation in cognitive science. arXiv preprint arXiv:2510.03311.

Peirce, C.S. (1931-1958). *Collected papers of Charles Sanders Peirce*. 8 vols. C. Hartshorne, P. Weiss, and A. Burks, (Eds.). Cambridge, MA: Harvard University Press.

Quilty-Dunn, J. (2016). Iconicity and the format of perception. *Journal of Consciousness Studies*, 23, 255-263.

Ragni, M., & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review*, 120, 561.

Ragni, M., Sonntag, T., & Johnson-Laird, P. N. (2016). Spatial conditionals and illusory inferences. *Journal of Cognitive Psychology*, 28, 348-365.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Regier, T., & Carlson, L. A. (2001). Grounding spatial language in perception: an empirical and computational investigation. *Journal of Experimental Psychology: General*, 130, 273.

Roth, J. C., & Franconeri, S. L. (2012). Asymmetric coding of categorical spatial relations in both language and vision. *Frontiers in Psychology*, 3.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking.* Cambridge, MA: MIT Press.

Sapkota, R., Flores-Calero, M., Qureshi, R., Badgujar, C., Nepal, U., Poulose, A., ... & Karkee, M. (2025). YOLO advances to its genesis: A decadal and comprehensive review of the You Only Look Once (YOLO) series. *Artificial Intelligence Review*, 58, 274.

Tran, T., Khemlani, S., & Trafton, J. G. (2025). Vision language models have difficulty recognizing virtual objects. arXiv preprint arXiv:2505.10453.

Tversky, B. (2005). Visuospatial reasoning. *Cambridge Handbook of Thinking and Reasoning*, 209-240.

Vukovic, N., & Williams, J. N. (2015). Individual differences in spatial cognition influence mental simulation of language. *Cognition*, 142.

Yu, S., Chen, Y., Ju, H., Jia, L., Zhang, F., Huang, S., ... & Lu, H. (2025). How far are vlms from visual spatial intelligence? a benchmark-driven perspective. arXiv preprint arXiv:2509.18905.

Yuan, L., Uttal, D., & Franconeri, S. (2016). Are categorical spatial relations encoded by shifting visual attention between objects?. *PLOS ONE*, 11, e0163141.

Zhang, J., Huang, J., Jin, S., & Lu, S. (2024). Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46, 5625-5644.