Chapter X –FROM PIXEL TO PERCEPT: TRACKING STATIC AND DYNAMIC SPATIAL RELATIONS IN VIDEO

Sangeet S. Khemlani
Navy Center for Applied Research in Artificial Intelligence
US Naval Research Laboratory
Washington, DC, USA

X.1 GENERAL BACKGROUND

In conventional intelligence, surveillance, and reconnaissance (ISR) operations, human analysts are often overwhelmed by the sheer volume of video footage that needs to be tracked and analyzed. NATO, US Navy Intelligence, and US Department of Defense stakeholders have expressed an overarching vision for integrated interactions amongst automated analytical systems and operators, particularly for automated data fusion and interpretation in the pursuit of command, control, communications, computers (C4) and ISR services (see the NRE Addendum to the Naval R&D Framework and the National AI R&D Strategic Plan, p. 8). These automated systems must understand, anticipate, and respond to demands fast enough to facilitate decision-making, particularly when they can exploit computational resources to analyze data in ways that would be prohibitive for human teammates. Time-consuming tasks, such as searching through surveillance data for specific details in a scene, can be automated to serve analytical needs. Machine learning (ML) methods can be used to train object detectors for Navy-relevant tasks such as passive tagging and automated spatiotemporal description of vehicles from surveillance data. Object recognition networks operate by processing image data to classify objects, assign them category labels, and report their locations in a scene. Recent systems are efficient enough to detect hundreds of objects in real-time with minimal computational overhead [6, 21].

However, state-of-the-art deep learning classifiers are limited in their ability to produce human-legible descriptions of scenes. Any given scene – of, e.g., vehicles in a parking lot – can be described in innumerable ways, and humans exhibit biases and preferences for a small subset of those descriptions, namely those that are concise, coherent, and relevant to task demands. Many of these human biases remain opaque to ML engineers, and the Navy and DoD have not investigated the cognitive constraints relevant to analytical information processing. Humans exhibit systematic strategies and preferences when processing scene descriptions, and automated systems must produce outputs that adhere to those biases to serve as effective teammates [3]. As a consequence, Naval Intelligence and the DoD maintain vast amounts of unlabeled and underutilized data.

To address this untapped resource, we developed a method for tracking static and dynamic spatiotemporal relations on Navy-relevant datasets. Our approach converts such data into meaningful, understandable, and actionable information that analysts can use to make decisions. From a research and functionality perspective, the research produced computational processing architectures that permit high-level relational analysis, both for static relations as well as for dynamic relations, such as a pursuit between one entity and another. A fundamental feature of the approach is that the architecture is designed to be cognitively plausible, i.e., subject to the same memory, attentional, and inferential limitations that intelligence analysts face.

The research highlighted here is a novel and unique departure from traditional analyses in two ways: 1) it is based on computations on sparse iconic spatiotemporal simulations, which are efficient modes of representing spatiotemporal relations [8]; and 2) the relevant spatiotemporal relations concerns those that are core to human cognition in the context of egocentric, point-of-view perspectives, unlike the relations described by orthodox spatial and temporal calculi. Sparse iconic spatiotemporal simulations demand unique processing capabilities that are nevertheless computationally tractable. In essence, iconic simulations are akin to small-scale diagrams of the scene; when the conventions for creating the simulations are known, they can be computed rapidly and stored in memory efficiently. We have demonstrated that these computational cognitive models are relevant

for small databases of limited test videos using everyday objects, but also that they operate on real-world scenes as well, particularly those that depict salient spatial arrangements of landmarks in 3D environments.

X.1.1 Automating anticipatory intelligence

A computational model designed to track static and dynamic spatial relations in video could fundamentally transform intelligence analysis by automating the synthesis of complex visual data into human-legible summaries. In ISR operations, human analysts are often overwhelmed by the sheer volume of video footage in need of analysis. A cognitively grounded model, such as one utilizing iconic spatiotemporal simulations, can identify critical spatial arrangements—e.g., relative positions of vessels or vehicles—and track dynamic behaviors like pursuit detection. By converting raw pixels into qualitative relational tags (e.g., "Vehicle A is following Vehicle B"), the system reduces the cognitive load on analysts and operators, allowing them to focus on high-level decision-making rather than manual target tracking.

Furthermore, these models are essential for the advancement of *anticipatory intelligence*, where the goal is to predict future actions based on current spatial trajectories and intent. By modeling the cognitive biases and attentional strategies that humans use to perceive animacy and pursuit, an AI system can track the latent goals of actors within a scene. For instance, distinguishing between a vehicle that is simply “following” another versus one that is “pursuing” with the intent to intercept allows for a more nuanced threat assessment. Integrating these computational cognitive models into autonomous systems enables them to anticipate analytical needs and provide predictive alerts, ensuring that human-machine teams can respond to emerging threats faster than manual analysis of streaming data would permit.

X.2 OBJECTIVE

Static spatial relations are those in that occur when two objects are in fixed locations relative to one another, such as when one vessel is *in the same place* as another. *Dynamic* spatial relations are those that occur when two objects are in motion relative to one another, such as when one vessel is *in pursuit* of another. This research project’s objective was to develop a system of computing with static and dynamic spatial relations from streaming and video data sources. It represented a novel and unique departure from traditional analyses in three ways: 1) it was based on computations on sparse iconic spatiotemporal simulations, which are efficient modes of representing spatial relations; 2) the relevant spatial relations under investigation were those that are core to human cognition in the context of egocentric, point-of-view perspectives, unlike the relations described by previous spatial and temporal calculi; and 3) computation was guided by models of human visual attention and reasoning, which focused processing on task-relevant object and relations while ignoring irrelevant details. Sparse iconic spatiotemporal simulations demand unique processing capabilities that permit novel forms of computationally tractability. In essence, iconic simulations are akin to small-scale “diagrams” of the scene; when the conventions for creating these diagrams are known, they can be interpreted rapidly. To cope with large numbers of objects in a scene, human-like attention mechanisms focused processing on the objects most likely to be relevant to a task communicated by those teammates.

This project’s ultimate goal was to design, develop, and test a novel computational system for tracking spatial relations from perceptual data using an integration of computational cognitive modeling, attentional modeling, and object recognition techniques. The project helped establish the first and only existing general purpose computational cognitive model of pursuit detection, as well as new insights into the high-level summaries of human behaviors as they reasoned about dynamic spatial relations. These results can be used to facilitate communication between personnel and machine teammates in fast-moving scenarios.

X.3 EXPERIMENTS

X.3.1 Experiments that benchmark pursuit detection behavior

To perceive dynamic spatial relations, people must track the objects over time and compare their locations, such as when an individual tracks one object in pursuit of another. The perception of pursuit and other dynamic relations is critical to our understanding of the behavior of those around us. A large body of research has explored the features that support and hamper pursuit detection. Recent pursuit studies used tasks designed to probe the perception of pursuit in dynamic stimuli [17-19]. These studies sought to test various cues that impact how people identify a pursuer object and a target object: for instance, set size, that is the number of objects in a display [17], the direction that the pursuer faces [5], the degree to which the pursuer deviates from the most direct path to the target [5, 17], and the distance between the pursuer and target [18, 19] can all affect the online detection of pursuit.

However, the tasks used in these studies were limited in their ability to highlight the factors contributing to success or failure in pursuit detection. They did not isolate the pursuit detection task from other cognitively demanding tasks – in a typical task, participants looked for pursuit among a field of identical, moving objects, resulting in potential difficulties with tracking the objects over time; and they did not provide a full picture of the error patterns produced by people; when participants incorrectly responded that pursuit was occurring, they were never asked to indicate which objects they believed were the pursuer and the target. In addition, many studies showed a complete video before asking participants to respond, limiting their ability to determine at what point in the video participants recognized that pursuit was occurring.

We therefore ran novel experiments designed to provide a better understanding of the factors that support or impair perception of pursuit. The results from the behavioral experiment are used to evaluate computational strategies and stopping rules and arrive and derive hypotheses about how people complete the task.

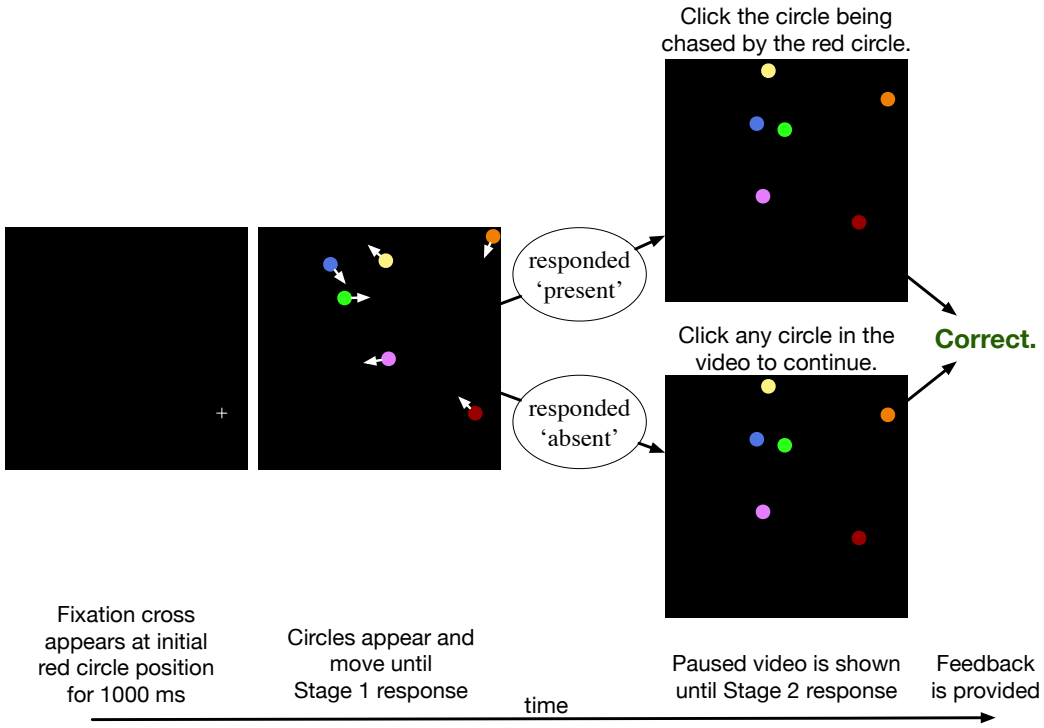

Figure X-1: Typical trial depicted to participants in experiments described in [15] and [16].

X.3.1.1 Experimental design, materials, and manipulations

Experimental videos and dynamics were developed using custom JavaScript and HTML code within the nodus-ponens package [10]. Figure X-1 provides a schematic example of the videos participants saw in a trial. A trial began with a black square and a fixation cross to highlight where the pursuer circle would be in a video that would begin 1000 ms later. The video showed either two or six uniquely colored circles, with colors taken from a colorblind-friendly palette.

At the video's start, circles were assigned a random initial direction and trajectory, such that no two circles overlapped at the video's start, and then they moved at a fixed speed for 22 seconds. Circles bounced off edges of the square but were allowed to pass through each other when their trajectories intersected. To determine which circle should appear in front of the other during these intersections, circles were assigned to random layers, such that a circle on a higher layer would appear in front of the other. The one exception was the red pursuer circle, which was always on the top layer and thus always appeared in front of other circles. The red circle also differed in its motion pattern – it updated its direction every 50 ms, so that it moved towards the target circle in the pursuit-present condition, or towards an invisible circle not shown to participants in the pursuit-absent condition. Finally, the videos were constrained such that (a) there was no point at which the target circle was completely covered by a distractor circle, and (b) there was no point at which the red pursuer contacted the target circle.

Four factors were manipulated in various studies: set size (2 or 6 objects), chasing condition (whether a pursuit was present or absent), chasing subtlety (the degree to which the chaser deviates from the most direct path to its target, i.e., 0°, 30° or 60°), and target color (one of five possible colors). Three videos were generated for each combination of conditions, producing 180 total videos. Chase-absent trials were generated the same as chase-present trials, except that the target was an invisible circle. Thus, the red circle's motion patterns did not differ depending on whether it was chasing a visible circle or not.

X.3.1.2 Experimental results

We subjected data to ANOVAs that analyzed accuracies at different stage of assessment (see Figure X-1) and also reaction times. Participants were largely accurate for each individual condition, i.e., they were accurate on 94% of the trials in Kon et al. [13]. Analyses of results suggest that the greater the number of objects in a scene, the longer it takes to detect whether there is a pursuit, which replicates previous findings [17]. Likewise, as the number of objects in a scene increases, participants are more likely to misjudge whether there is a pursuit and misidentify the target circle (Figure X-2). These results establish benchmark patterns in pursuit detection: reaction times increase as set size increases, and reaction times tend to be greater for absent than present trials.

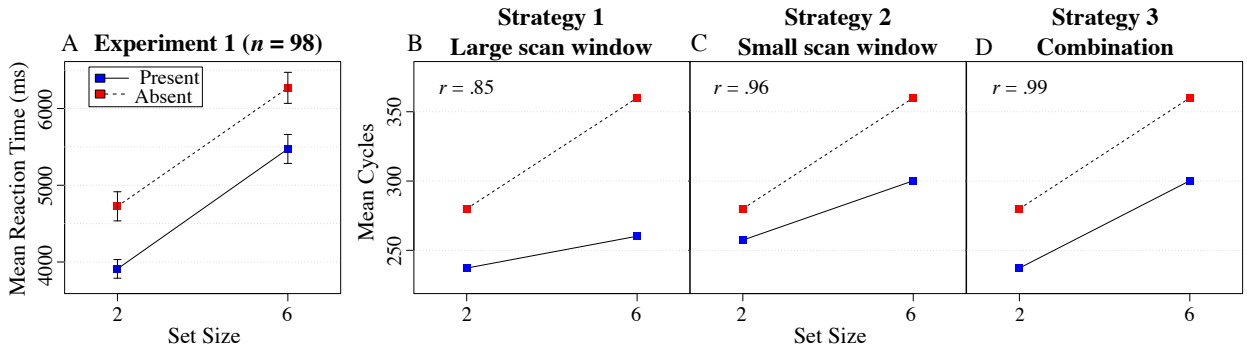

Figure X-2: *Panel A:* Human data from a typical experiment on chase detection behavior [15]; *Panels B, C, and D:* Variations of computational modeling results that show fits to data using a large scan window (A), a small scan window (B), and a combination of strategies (C).

A key strength of present experimental paradigm is that it can provide detailed information on human error patterns. We focus here on the errors for one particular video (hereafter called 'video 1'). This was a pursuit-absent trial where a disproportionate number of participants incorrectly responded 'present' in Stage 1. In fact, this video received 45% of the total incorrect `present' responses across all videos in Experiment 1 [15]. The Stage 2 responses to this video, in which participants indicated which circle had been targeted by a pursuer, are depicted in the left graph of Figure X-3. As the graph indicates, nearly all participants believed the pursuer had targeted the green circle. On reviewing video 1, which had 6 circles, we believe that the red circle could be interpreted as tracking the green circle from about the 4.5 second mark until the 7 second mark, after which time it heads away from the green circle. Notably, the mean Stage 1 reaction time for these incorrect trials was *M*=5.73, *SD*=2.06, during that window of time when the red circled appeared to be pursuing the green circle, whereas the mean time for correct responses was *M*=7.29, *SD*=2.77, after it may have become apparent that there was no pursuit.

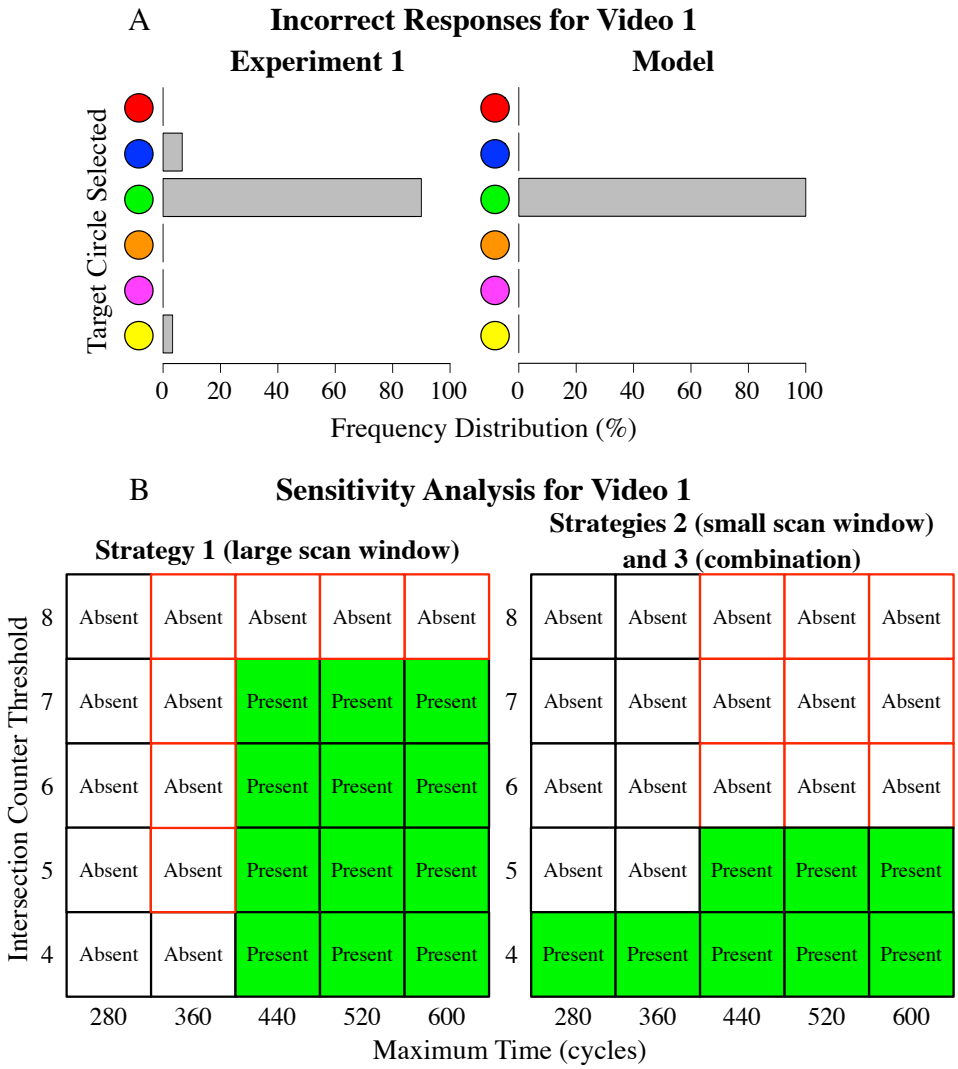

Figure X-3: *Panel A:* Percent of trials where a particular colored circle was selected as the target after observers incorrectly responded 'present' on a difficult pursuit-absent video ('video 1'). *Panel B:* Sensitivity analysis showing, for each strategy, what pairs of parameter values resulted in correct 'Absent' responses or incorrect 'Present' responses for video 1. The shaded green boxes indicate that the model always selected the green circle as the target after an incorrect 'Present' response.

As video 1 shows, systematic errors in pursuit perception may occur at various points of a sequence of movements. If errors are robust, then a cognitive model of pursuit detection should be able to simulate them. We turn to the development of a computational model that explains both benchmark patterns of pursuit detection as well as errors.

X.3.2 Experiments that test the semantics of dynamic relations

Everyday language contains "tracking" verbs such as pursue, chase, and follow, which share conceptual similarities: they are all action verbs that describe a dynamic spatial relation between two entities, as in "the cat chased the mouse". What distinguishes them from one another? We developed a conceptual theory that builds on the proposal that action verbs are abstract, modality-independent representations: they help humans construct mental simulations rather than activate sensorimotor information. Modality-independent representations can encode for concepts such as desire and intentionality, and this information can modulate how the mind constructs dynamic and kinematic mental simulations to represent tracking verbs. Some tracking verbs are more intentional than others. For instance, chase and pursue convey a pursuer's intention of catching up to and intercepting its target, whereas follow conveys no such intention. We conducted eight experiments that tested whether intentional tracking verbs yield conceptually faster motion [13, 14]. The studies presented participants with imagery of one car chasing another along a straight road. For some experiments, participants estimated the distance that the pursued car or both cars would travel 3 seconds into the future. For others, participants estimated the rate of one or both cars. Other experiments tasked participants with estimating the duration of the pursuit. Altogether, the studies validated the hypothesis that chase and pursue describe faster motion, i.e., participants reliably estimated longer distances for descriptions that included those verbs.

X.3.2.1 Experimental design, materials, and manipulations

Experiments presented participants with problems concerning the movements (or lack thereof) of two colored cars on a straight road (see Figure X-4). On each problem, participants evaluated the cars relative to a sentence description of the scene, which appeared above the car image as shown in Figure X-4. Some of the problems (i.e., experimental problems) described the scene using a tracking verb (e.g., "the yellow car is pursuing the blue car"), while other problems were controls (e.g., "the yellow car is parked and the blue car is moving forward slowly"). The experimental conditions were the same across all experiments. Additionally, the control conditions were the same for all experiments except for those that measured estimated duration.

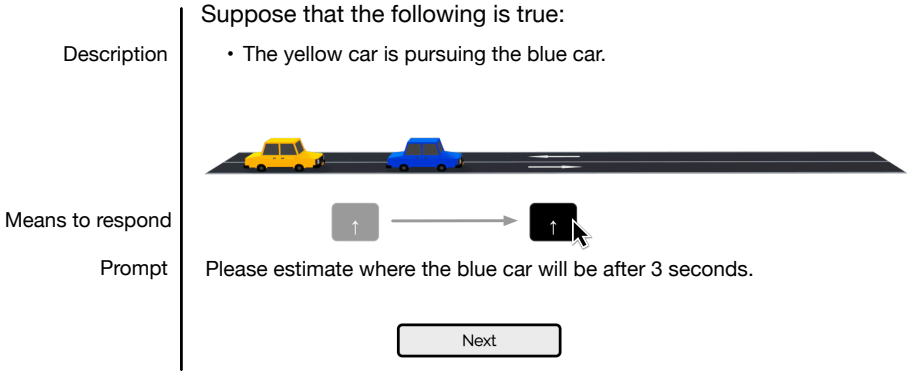

Figure X-4: An example trial shown to the right of the vertical bar with its components labeled for ease of reference on the left for experiments described in [13, 14]. On each trial participants saw two cars along a road below a description. Below the image a prompt and a means of responding was provided, both of which varied depending on the experiment. The example shows the slider task used in Kon and Khemlani Experiment 1 [13], where, on each problem, a slider handle appeared between the two cars. Participants used their mouse to drag it to a location that satisfied their intuitions about how far the specified car would move. For all experiments, participants registered their response and moved to the next trial by clicking the 'Next' button.

Participants carried out each experimental problem twice, and the problems appeared in a different random order for each participant. Each car appeared at the same position on the road, but their color varied across trials; the experiment assigned a separate pair of colors (e.g., blue-yellow, green-white, and so on) for the cars on each trial. Figure X-4 shows an example trial where the left car is yellow and the right car is blue. Specific color values were drawn from a colorblind-friendly palette. As shown in Figure X-4 with an example trial from Experiment 1 in [13], the means of responding and the prompt were shown below the image. These aspects varied depending on the task of the experiment.

X.3.2.2 Experimental results

Across various experiments, participants estimated larger distances of travel for experimental than control problems. In Experiment 1 of [13], which we use as a benchmark and foundation for further studies, participants estimated the mean slider value to be 72.4 for experimental vs. 52.1 for control trials, and there was a statistically reliable difference between these estimates ($p < .001$, Cliff's $\delta = .60$). Their responses to control problems yielded a significant trend: lower estimates for parked controls (mean slider value = 42.2), middling estimates for slowly (mean slider value = 50.5), and higher estimates for quickly (mean slider value = 63.7; trend test, $p < .001$). The results likewise demonstrate systematic comprehension of the task. Participants made larger estimates for *pursue* (mean slider value = 74.1) and *chase* (mean slider value = 74.9) than for *follow* (mean slider value = 68.1). Pairwise analyses showed reliable differences between *pursue* and *follow* ($p < .001$; Cliff's $\delta = .20$) and between *chase* and *follow* ($p < .001$, Cliff's $\delta = .23$). There were no reliable differences in estimates between *pursue* and *chase* ($p = .18$, Cliff's $\delta = .04$). The results corroborated the hypotheses driven by the theory we described.

Other experiments replicated these general trends. For instance, Experiment 2 in [14] used the exact same design and materials as the experiment whose results were described above, except it asked participants to make estimates by selecting from a series of options provided on a screen (not a slider). Figure X-5 shows the distance estimates that participants made in Experiment 2.

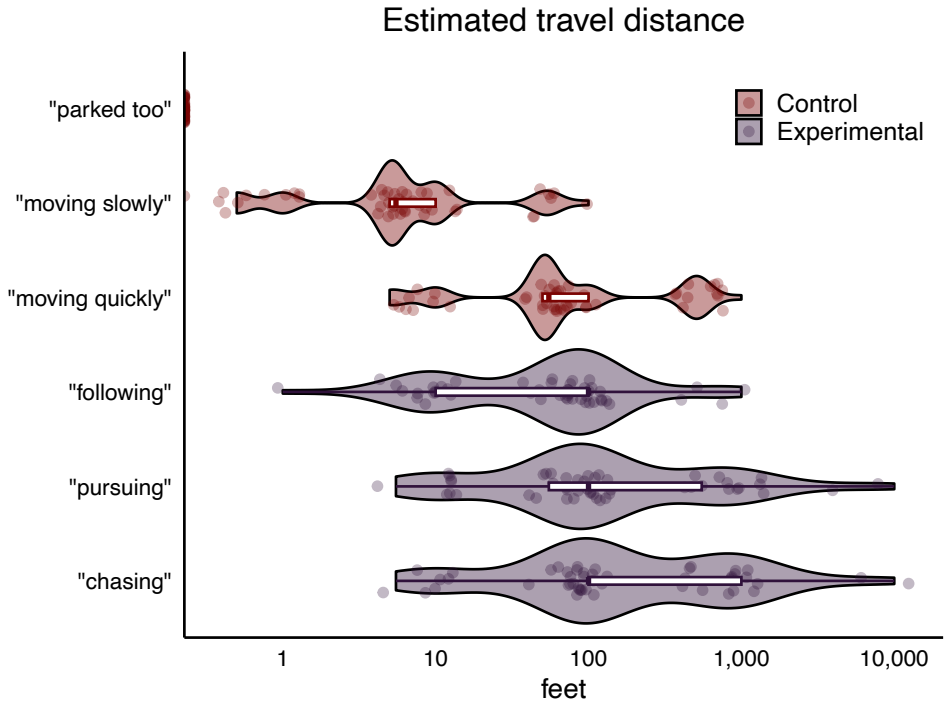

Figure X-5: Violin and box plots of estimates of travel distance in Experiment 2 [14]. Control conditions are pink, and the experimental conditions are purple. The *x*-axis indicates estimated travel distance in feet. Each dot represents the mean response of a participant for the condition.

In summary, the results of these studies show that human reasoners can both visualize and conceptualize dynamic relations such as *chase* events, *pursuit* events, and *following* events. Several cues help you to perceive chasing behavior, such as the manner in which a pursuer moves towards its target. Indeed, simulations of perceptual events may help individuals discuss and describe chases, which in turn they can simulate by constructing and animating spatial mental models. The results reveal preliminary evidence that individuals distinguish between chasing, following, and pursuit by the ways in which they mentally animate those models.

X.4 ANALYSIS OF SPATIAL RELATION TRACKING IN VISION LANGUAGE MODELS

Vision language models (VLMs) are designed to extract relevant visuospatial information from images. Some research suggests that VLMs can exhibit humanlike scene understanding, while other investigations reveal difficulties in their ability to process relational information. To achieve widespread applicability, VLMs must perform reliably, yielding comparable competence across a wide variety of related tasks. We sought to test how reliable these architectures are at engaging in trivial spatial cognition, e.g., recognizing whether one object is left of another in an uncluttered scene. We developed a benchmark dataset – TableTest – whose images depict 3D scenes of objects arranged on a table, and used it to evaluate state-of-the-art VLMs [12, 22]. Results show that performance could be degraded by minor variations of prompts that use logically equivalent descriptions. These analyses suggest limitations in how VLMs may reason about spatial relations in real-world applications. They also reveal novel opportunities for bolstering image caption corpora for more efficient training and testing.

X.4.1 Benchmarking spatial relation recognition in vision-language models

We developed TableTest, a synthetic dataset designed to test spatial relation recognition and reasoning in VLMs. Each image in the dataset depicts one or more objects arranged on a table. We selected 64 objects from the Objaverse dataset of annotated 3D objects based on the following criteria:

- Selected objects had limited association with one another (e.g., blender and screwdriver) instead of all coming from a given environment (kitchen appliances).
- All objects were small enough to sensibly rest on one portion of a table, but nevertheless varied in size, material, color, and texture.
- Objects included both natural items (tomato, banana) as well as artifact kinds (telephone, lamp).

We hand-scaled these objects to ensure plausible relative object sizes, and used Blender 3D to ensure uniform table placement, such that the center point of each object was placed 200px away from the center point of the table in the left or right directions. TableTest consists of all possible 2- and 3-object configurations of the 64 objects, yielding a dataset of 4,032 2-object images and ~250K 3-object images. We used TableTest to conduct evaluation tests of spatial relation recognition behavior in readily available VLMs.

Spatial relation recognition behavior should be accurate and consistent across different kinds of verbal prompts, such as those provided in the introductory paragraph above. To test the capacity of VLM architectures to cope with relation recognition, we subjected common VLMs to a battery of prompts that varied in task and format. We included VLM architectures for evaluation based on three criteria: they were recently released (post-2022), freely available and well documented, and capable of single-shot identification of the 64 objects in TableTest without additional training. We identified three architectures that matched those criteria: Idefics2 (8B parameters), InstructBlip-Vicuna (7B parameters), which builds atop the BLIP architecture, and Llama 3.2 (11B parameters). Initial testing verified that all three VLMs identified the 64 single-object images in the dataset at or above 96% accuracy.

X.4.2 Omnibus results of benchmarking studies

Evaluation of spatial relation recognition consisted of two phases. The first phase examined multimodal input: we subjected each two-object image in TableTest to eight separate prompts that queried for spatial relations depicted in the image (see [12] for detailed prompt descriptions). A second comparison phase used text-only translations of those same TableTest images to run analogous queries. Prompts (and image descriptions) were minimal in nature to permit systematic comparison. For example, Prompt 4 asked for yes/no answers to questions such as "Is the blender to the left of the cupcake?" An advantage of this approach was that each prompt could be used to identify relations from either multimodal or text-only input.
We likewise devised surface-level structural variations of the different prompts to test for uniformity in VLM responses. Each of these variations should have no effect on VLM relational recognition. For example, we asked 4 separate versions of Prompt 1 by varying a) whether the prompt used *right* or *left* in its description, and whether the image depicted the relation or not.

As Figure X-6 shows, VLM performance across different prompts reveals unreliable performance for both multimodal and text-only evaluations. Adequate performance should be > 90% accuracy for every prompt and prompt variation under investigation. Overall accuracies across 8 different prompts ranged from .12 to 1.00. Ranges for the different VLMs were as follows: BLIP [.12, .96]; Idefics2 [.20, 1.00]; Llama3 [.12, .98]. VLMs reached adequate performance on 3 of the 8 prompts used in multimodal contexts and 7 of 8 prompts used in text-only contexts. For multimodal contexts, Llama3 outperformed other models on 5 prompts; Idefics2 outperformed other models on 3 prompts; and BLIP consistently underperformed relative to other models. For text-only contexts, Idefics2 outperformed other models on 6 of 8 prompts; Llama3 outperformed other models on 2 of 8 prompts; and BLIP underperformed.

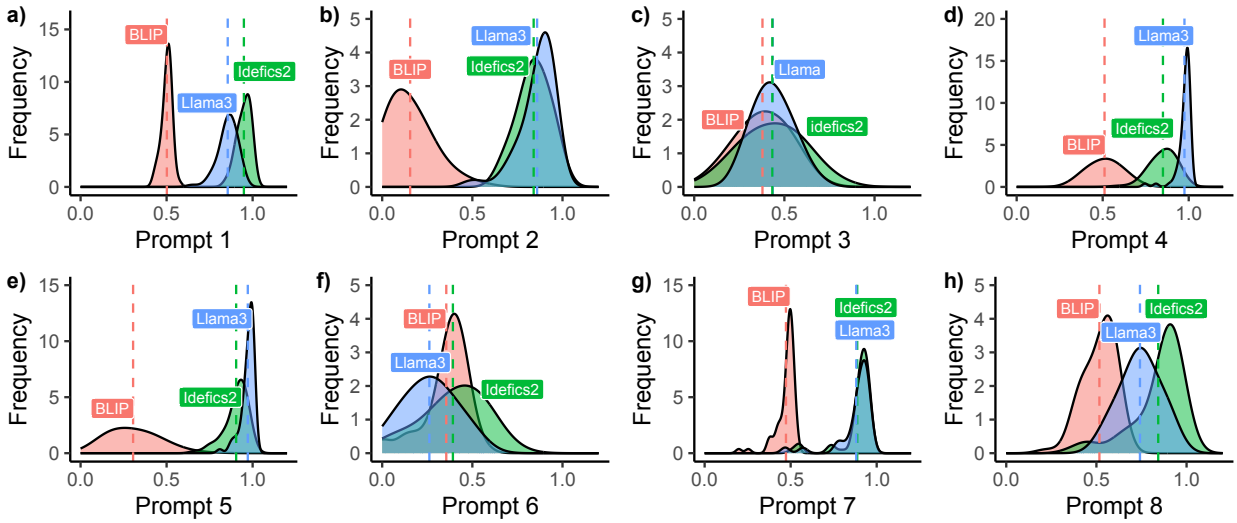

Figure X-6: Proportion accuracy from multimodal evaluations of VLM performance on spatial recognition prompts (see Khemlani et al., 2025) in which each prompt was paired with an image from TableTest. Dashed lines in each panel depict overall accuracies of each VLM and density plots depict performance distributions across TableTest's 64 objects. Humanlike performance anticipated to be at ceiling (accuracy = 1.0).

In sum, none of the models we tested revealed consistent and reliable spatial relation comprehension, either from visual imagery or text. These results suggest limitations in how reliable such models are at higher-order spatial reasoning, which depends on the reliable recognition and processing of spatial relations. If advances yield VLM architectures the exhibit higher reliability, designers will need to engage in investigations such as the one we conducted with TableTest above, i.e., those that test across multiple prompts using imagery in which the spatial properties of the objects are known in advance.

X.5 COMPUTATIONAL COGNITIVE MODELING

X.5.1 Computational modeling of pursuit detection behavior

We developed a cognitive model for the chase detection task [15, 16], to aid in exploring problem-solving strategies and stimulus factors that may influence performance. The core claim of this model was that chase detection, like other demanding visual tasks, depends on strategically projecting spatial attention onto a visual scene. To determine whether an object is a pursuer, the model tracks the object, determines its motion trajectory, and then projects spatial attention along that trajectory to identify another object that may be the target of pursuit. This process is engaged repeatedly, and if a prospective pursuer is consistently moving towards the same prospective target, then it is likely that a chase is occurring.

We evaluated our model on a novel chase detection task, in which the possible pursuer was always a red circle and the possible targets were circles of unique colors (see Figure X-7 below). This color-coding simplified detection and tracking in a way that isolated specific factors that contribute to chase detection. The task also utilized a two-stage design in which participants first pressed a button to indicate whether a chase was occurring (Stage 1) and then indicated the circle being chased (Stage 2), allowing for measures of response time and accuracy.

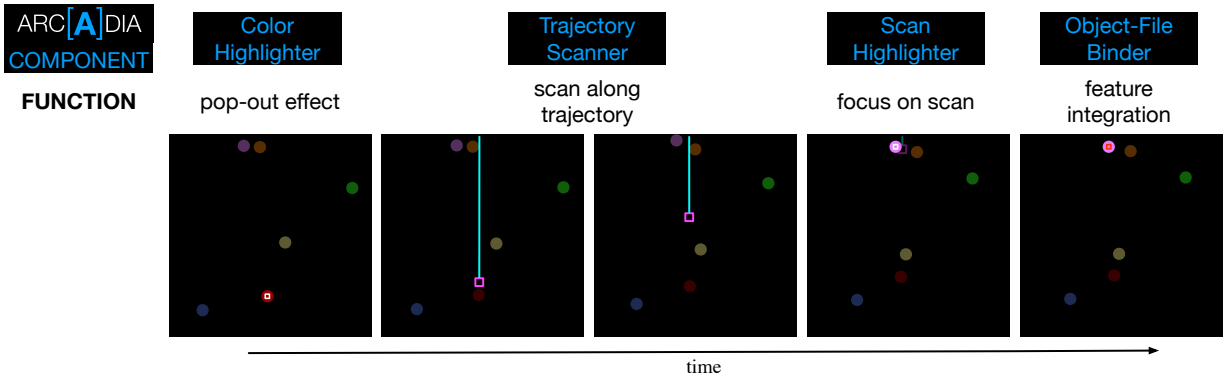

Figure X-7: Model developed in the ARCADIA cognitive system for attentional processing, which processes moving objects and engages processes that track potential targets of pursuit by a) popping out objects of relevant interest (i.e., the pursuer object); b) scanning along the trajectory of the pursuer object's motion; c) selecting objects that are detected during the trajectory scan; and d) binding features of those objects together for purposes of tracking; the model is described further in Kon et al. (2024, 2025).

After gathering human data on the task, we presented the same trial videos to the model and found an overall close fit to the human data, thereby providing support for the model (see Figures X-2 and X-3 above).

X.5.2 Computational modeling of embodied spatial reasoning for operational maintenance tasks

Despite being data-, power-, and resource-greedy, many trillion-parameter language models trained on internet-scale corpora can't draw trivial deductive conclusions at human competence [9]. These limitations prevent many advances in automating real-time, interactive, decision critical tasks. Our computational modeling approach appeals to computations based on iconic possibilities to maximize cognitively plausibility; but preliminary investigations show how possibilities can serve as optimal ways of packaging abstract, symbolic, relational information to permit fast, reliable AI. Figure X-8 shows how: consider the 12 mutually consistent spatial relations in (8a). Could these relations be encoded in a single representation? Logic-based AI systems eschew such integration, and instead treat them as separate entries in a knowledge base [11]. It's possible to depict a scene representing those relations in an image (8b); but images can be

taxing to analyze because they encode irrelevant information, such as backgrounds and shadows. Machine learning techniques can extract relevant information about the objects and their locations in the image (bounding boxes in 8b), but this approach relies on treating image patches and detected objects as vectors in an embedding space (8c), and inferences on these vectors are themselves computationally expensive vector transformations. In contrast to all these alternatives, a spatial possibility (8d) can sparsely encode all the abstract information in the 12 relations. This sparse representation dispenses with imagistic information to reduce the computational burden of scanning and comparing its structure against a spatial semantics. Spatial possibilities place low demands on data and power, and are therefore feasible for real-time inference applications.

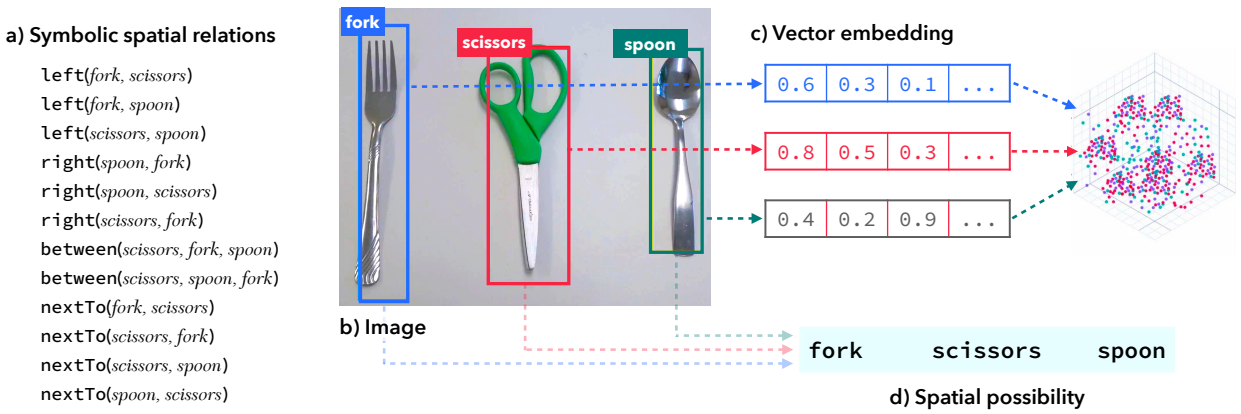

Figure X-8: 12 symbolic static spatial relations that are mutually consistent with one another (a) as well as different ways of constructing unified representations of them (b, c, d). Images (b) represent symbolic relations and irrelevant details. Machine learning systems can construct vector embeddings of relations and imagery (c), but those embeddings are expensive to process. Spatial possibilities are iconic representations that minimize processing demands (d).

We developed a computational cognitive theory of applying dimensionality reduction techniques depicted in Figure X-8 to object recognition ML systems for generating discrete qualitative relation tags for Navy relevant maintenance environments, i.e., the maintenance of electrical panels. This approach works by converting continuous and probabilistic output from an object detection convolutional neural network (CNN) to compute a vector of object labels and a vector of their bounding-box dimensions (Jiang et al., 2022). The resulting representation serves as a sparse iconic spatial simulation of the scene [1, 2, 7, 8], which can then be used to extract a small subset of relevant spatiotemporal relations. The methodology adapt a cognitive spatial possibility semantics [20] and extends it to cope with binary and ternary relations for 2D scene processing.

Figure X-9 shows how this data analytic pipeline can be applied towards the interactive maintenance of an electrical panel. Object recognition finds relevant components (e.g., a terminal block and two relays) and their corresponding bounding boxes of the electrical panel (Figure X-9a). Those results are downsampled and their imagery discarded to create a spatial arrangement of components (akin to Figure X-8d). This downsample occurs in linear time, permitting efficient analysis of each individual frame of video. Individual frames are coalesced via a “heavy hitters” algorithm to reduce noise and construct a stable spatial possibility that mirrors the scene observed in the video. The resulting scene representation is then scanned to respond to natural language commands, such as “find whatever is to the left of the terminal block.” The system answers reliably and quickly. It does so because the engine of possibility at its core packages abstract relational information in iconic structures, which permits rapid inference.

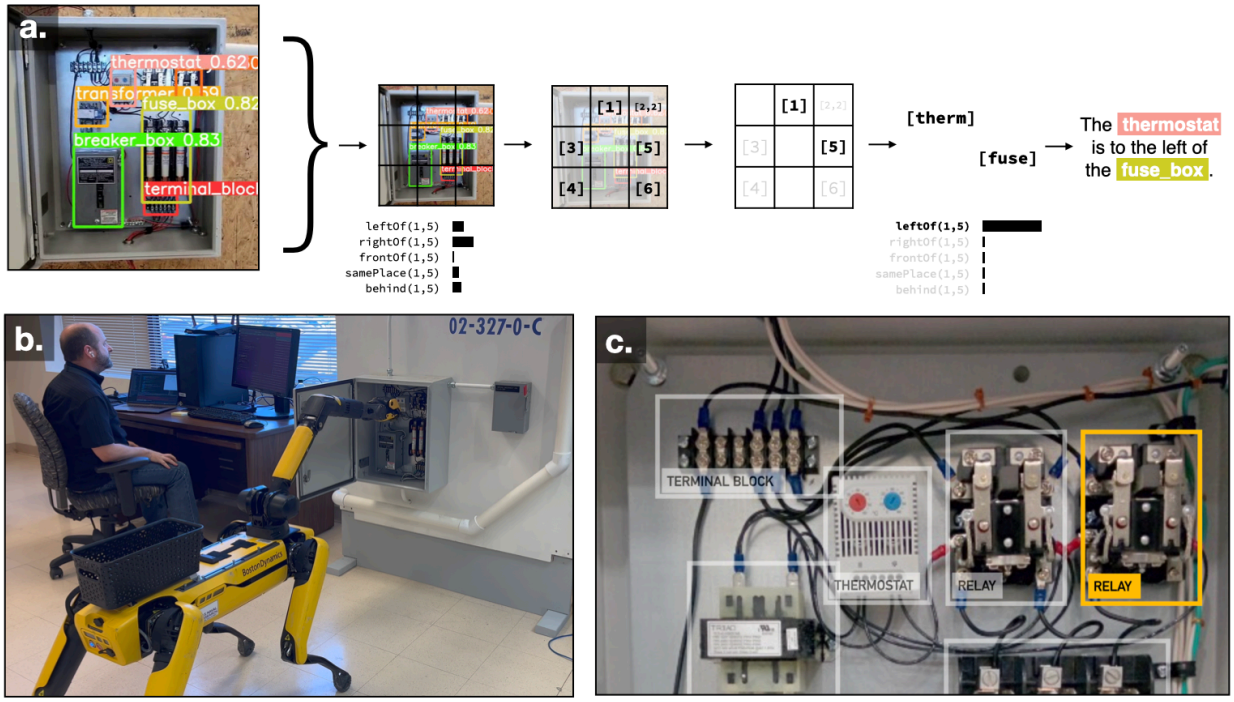

Figure X-8: The ASTEREX perceptual processing algorithm applies a dynamic grid downsample to reduce the complexity of constructing spatial possibilities (a). The algorithm tags each frame of data with qualitative spatial relations; frames can then be parsed and searched to identify regions of similar/dissimilar spatial arrangements. Because the algorithm can build a spatial possibility of a scene, it can also describe it using interpretable language to assist with human-machine teaming (b), as in the case of routine maintenance operations. Here it accurately disambiguates components in an electrical box (c) by interpreting a prompt to locate whatever is "directly to the right of the leftmost relay."

X.6 CONCLUSIONS

This project helped to realize novel computational frameworks for tracking and reasoning with static and dynamic spatial relations from visual data. It investigated those spatial relations of utmost and fundamental importance to human spatial reasoning, in anticipation of intelligence analysis needs in Navy relevant environments. It proceeded along four dimensions: empirical evaluations that served as benchmarks for computational modeling; investigations that served to test and evaluate a novel theoretical account of the meanings of dynamic relations; benchmarks of spatial relation recognition in state-of-the-art vision language models; and cognitive models of both static and dynamic relations, which are capable of analyzing and summarizing relational information from video data. The project therefore resulted in novel datasets, novel methodologies for tracking spatial relations, and novel data-backed algorithms that track relations in the same manner that humans do.

The approach can be improved and extended in several ways. First, many of the static and dynamic relations tracked concern binary relations, e.g., *A is to the left of B*. In principle, the same methodologies could permit efficient analysis of more complex relations, such as ternary relations, e.g., *C is between A and B*. Second, the ARCADIA model of tracking dynamic pursuit relations could be extended to explore pursuit detection on Navy relevant stimuli. Indeed, we have begun preparing materials – 3D test videos – to analyze pursuit between two agents in a crowded 3D urban environment, and whether and how our system can be extended to cope with such interactions. Third, the system we developed assumes a fixed, egocentric, first-person point of view (POV) against which spatial relations are tracked and analyzed. But many methodologies could be extended to top-down POVs. Top-down POV analysis may be most relevant for the Navy

Intelligence Community and geospatial intelligence analysts, who base reporting on satellite and drone reconnaissance. Fourth, we developed an embodied system of spatial reasoning for human-robot interaction assuming perfect performance from conventional object-recognition technologies. Such systems are never truly perfect: one limitation is that once they learn which objects to recognize, that learning is fixed. For novel objects, recognition systems must be trained and fine-tuned further. This presents an opportunity to develop methodologies that help mitigate the computational overhead of fine-tuning these systems.

In sum, this project successfully developed both theoretical and computationally grounded frameworks for the rapid analysis and sensemaking of spatial relations in video data. Its success has the potential to yield new insights, both into how humans perceive spatial relations as well as how they do so efficiently to build stable representations of their environment. By modeling these specific human capabilities in a manner that matches their performance, we can equip computational systems with the ability to perform rapid analysis of movements and activity in video data, and thereby bolster the intelligence capabilities to exploit video and surveillance corpora.

X.7 REFERENCES

[1] K. Alfred et al., “Mental models use common neural spatial structure for spatial and abstract content,” Communications Biology, vol. 3, p. 17, 2020.

[2] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum, “Simulation as an engine of physical scene understanding,” Proceedings of the National Academy of Sciences, vol. 110, pp. 18327–18332, 2013.

[3] V. Dentella, F. Günther, E. Murphy, G. Marcus, and E. Leivada, “Testing AI on language comprehension tasks reveals insensitivity to underlying meaning,” Scientific Reports, vol. 14, no. 1, p. 28083, 2024.

[4] T. Gao, C. L. Baker, N. Tang, H. Xu, and J. B. Tenenbaum, “The cognitive architecture of perceived animacy: Intention, attention, and memory,” Cognitive Science, vol. 43, no. 8, p. e12775, 2019.

[5] T. Gao, G. E. Newman, and B. J. Scholl, “The psychophysics of chasing: A case study in the perception of animacy,” Cognitive Psychology, vol. 59, pp. 154–179, 2009.

[6] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, “A review of YOLO algorithm developments,” Procedia Computer Science, vol. 199, pp. 1066–1073, 2022.

[7] P. N. Johnson-Laird and S. S. Khemlani, “Toward a unified theory of reasoning,” in Psychology of Learning and Motivation, vol. 59, pp. 1–42, Elsevier, 2013.

[8] P. N. Johnson-Laird and S. Khemlani, “Mental models and the algorithms of deduction,” in The Cambridge Handbook of Computational Cognitive Sciences, R. Sun, Ed. Cambridge University Press, 2023.

[9] S. Kambhampati, “Can large language models reason and plan?” Annals of the New York Academy of Sciences, vol. 1534, no. 1, pp. 15–18, 2024.

[10] S. Khemlani, “Nodus-ponens: A light, full-stack framework for running high-level reasoning and cognitive science experiments in Node.js,” npm package, 2024. [Online]. Available: https://www.npmjs.com/package/nodus-ponens

[11] S. Khemlani and P. N. Johnson-Laird, “Why machines don’t (yet) reason like people,” KI – Künstliche Intelligenz, vol. 33, pp. 219–228, 2019.

[12] S. Khemlani et al., “Vision language models are unreliable at trivial spatial cognition,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR) Workshops, 2025.

[13] M. Kon and S. Khemlani, “How spatial simulations distinguish ‘tracking’ verbs,” in Proc. 46th Annual Conf. Cognitive Science Society, Austin, TX: Cognitive Science Society, 2024.

[14] M. Kon and S. Khemlani, “Verbs that track,” manuscript under review at Cognition.

[15] M. Kon, S. Khemlani, and A. Lovett, “Measuring and modeling pursuit detection in dynamic visual scenes,” in Proc. 46th Annual Conf. Cognitive Science Society, Austin, TX: Cognitive Science Society, 2024.

[16] M. Kon, S. Khemlani, G. Francis, and A. Lovett, “Model predictions and implications for chasing subtlety,” in Proc. Int. Conf. Cognitive Modeling, 2025.

[17] H. S. Meyerhoff, M. Huff, and S. Schwan, “Linking perceptual animacy to attention: Evidence from the chasing detection paradigm,” Journal of Experimental Psychology: Human Perception and Performance, vol. 39, pp. 1003–1015, 2013.

[18] H. S. Meyerhoff, S. Schwan, and M. Huff, “Perceptual animacy: Visual search for chasing objects among distractors,” Journal of Experimental Psychology: Human Perception and Performance, vol. 40, pp. 702–717, 2014.

[19] H. S. Meyerhoff, S. Schwan, and M. Huff, “Interobject spacing explains the attentional bias toward interacting objects,” Psychonomic Bulletin & Review, vol. 21, pp. 412–417, 2014.

[20] M. Ragni and M. Knauff, “A theory and a computational model of spatial reasoning with preferred mental models,” Psychological Review, vol. 120, no. 3, p. 561, 2013.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 779–788, 2016.

[22] T. Tran, S. Khemlani, and J. G. Trafton, “Vision language models have difficulty recognizing virtual objects,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR) Workshops, 2025.