Step, Switch, Repeat: How Mental Simulation Predicts Collisions

Hanbei Zhou (hanbeizhou@g.harvard.edu)
Department of Psychology, 52 Oxford Street, Cambridge, MA 02138 USA

Halely Balaban (halelyb@openu.ac.il)
Department of Education and Psychology, The Open University of Israel, Ra'anana, Israel

Tomer Ullman (tullman@fas.harvard.edu)
Department of Psychology, 52 Oxford Street, Cambridge, MA 02138 USA

Abstract

Previous research suggests a capacity limit in dynamic mental imagery: people simulate the motion of only one object at a time (Balaban & Ullman, 2024). This raises an obvious question: how do people imagine interactions between multiple moving objects? Here, we used a novel paradigm to examine people's estimates of the timing and location of collision events that happened either between (1) two imagined moving entities, or (2) one imagined moving entity and a static patch. We propose and evaluate two mental simulation models: In the Parallel model, both entities are advanced simultaneously until collision. In the Switch model, the mind alternates between entities, updating one for several steps before switching to the other, and back again. We found that the Switch model better predicted participants' collision time estimates in the two-entity condition relative to the one-entity condition, whereas the Parallel model failed to capture this difference (Experiment 1 and Experiment 3). Moreover, the Switch model predicted systematic biases in reported collision positions that the Parallel model did not (Experiment 2 and Experiment 4). Together, these findings suggest that the mental simulation of interactions between multiple dynamic objects is better described by a Switch model, in which the mind alternates between entities, rather than simulating them in parallel.

Keywords: mental simulation; mental capacity; visual imagery, computational modeling

Introduction

People don't just see what's in front of them; they anticipate what is coming next. An unstable mug at the edge of the table will topple. A ball rolling behind a pile of books will emerge at the other end. Two people running blindly towards each other will collide. Such common-sense expectations about the dynamics of everyday life are fast, automatic, and develop early on (Spelke, 2022). A current line of research suggests that this kind of intuitive reasoning is based in part on running an approximate mental simulation (Battaglia et al., 2013; Hamrick et al., 2016; Ullman et al., 2017). The simulation approach suggests that the mind reconstructs a scene internally, and then unfolds its future states step-by-step.

Even if people use mental simulation, an internal scene cannot be a perfect replica of the real world, due to finite computational resources. Empirical evidence backs up this intuition, and suggests that mental simulation uses systematic approximations, and is deployed in a context-specific way (Bass et al., 2021; Sosa et al., 2025; Smith et al., 2023; Wang & Ullman, 2025). One constraint on mental simulation is a limit on the number of objects that can be simulated at once. Balaban & Ullman (2025) examined this limit by asking people

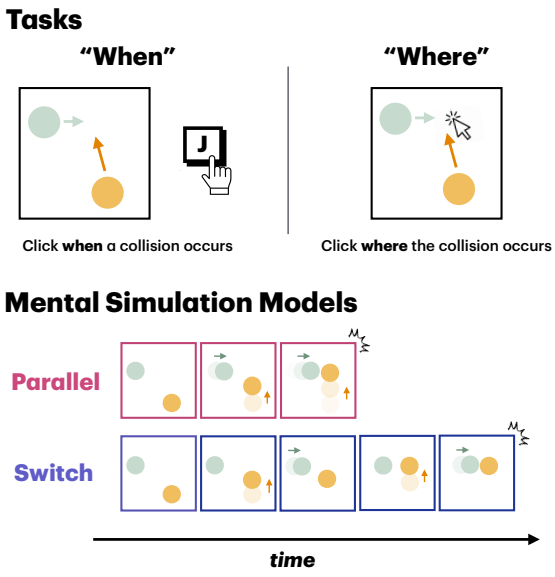

Figure 1: Overview of tasks in our studies and contrasting mental simulation models. People saw brief presentations of moving objects, and were asked to continue their motion in the imagination. People either indicated *when* or *where* a collision happened. We contrasted mental models according to which people either mentally moved both objects at the same time ('Parallel') or in an interleaved way ('Switch').

to imagine the continued motion of one or two falling balls and press a key when each hit the ground in their mind's eye. While estimates were accurate for a single ball, adding a second ball created a systematic delay, in line with the prediction of a 'serial' simulation model, according to which one object is simulated first, and only then is the second object simulated. This serial limit on updating dynamics in mental simulation is more strict than, and separate from the capacity limits of perceptual tracking, where multiple lines of research using the Multiple Object Tracking paradigm indicate people can track multiple objects at once (though the precise capacity limit and its nature remains debated Pylyshyn & Storm, 1988; Alvarez & Franconeri, 2007; Lovett et al., 2019).

A limit of one object in the updating of mental simulation presents an obvious puzzle: what if multiple moving bodies can interact? Consider a simple situation in which two balls

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

are on a collision course (Figure 1). If mental simulation is fully serial, the two objects will not hit one another in the imagination. This seems silly on the face of it. However, there is a way for the imagination to be serial and still lead to interactions and collision. Here, we propose a *Switch* model, by which the mind alternates the object being simulated: it updates one object for S steps before switching to the other, back and forth until some condition is met [1].

Across four experiments, we examined people's estimation of the **timing** and **position** of simulated collision events. We compared people's response when imagining two colliding entities, to their response when imaging one entity colliding with a static point. We compared people's behavior to the predictions of a Serial-Switch mental simulation model, and those of a Parallel model. More specifically, we examined people's estimation of the timing (Experiment 1) and position (Experiment 2) of the collision of objects in simple linear motion, and then examined the estimation of more complicated agent movement (Experiments 3 and 4). Across our studies, we found that the Switch model overall best accounted for participant behavior, though some aspects of people's behavior were not captured by either current Switch or Parallel models. Our results suggest that people's mental simulation does have a strict capacity limit when it comes to updating moving objects, but that people can bypass this limit to imagine interactions by interleaving their step-by-step simulation.

Experiment 1: "When" Task with Objects

We examined people's response times when imagining two simple objects moving towards each other on a collision path.

Method

Our first experiment included two conditions: a two-object condition, and a one-object condition. In the two-object condition, participants viewed two balls moving such that if their trajectories were continued in reality, they would collide. The balls paused mid-motion after 500 ms. In the one-object condition, we replaced one of the moving balls with a static patch, located at a collision point that was paired with the stimuli in the two-object condition. In both conditions, participants were asked to imagine the continued motion of the ball(s), and press a key when they estimated a collision occurred. We measured participants' estimated collision time using their response time (RT), which was the duration between the onset of the trial and the key press.

There were three collision *types* in this experiment: head-on, sideways, and chasing. In head-on, two balls move directly toward each other; in sideways, balls would collide at a 90° angle; and in chasing, both balls move in the same direction, with one faster than the other as if chasing it. The equivalent of the one-object control to the head-on and sideways collision meant randomly selecting one of the two balls to be moving, and the other placed statically at the collision point. In the one-object control to the chasing collision, the moving ball was always the 'chaser', otherwise a collision would not occur at the correct position. We used three different collision times: 1000 ms, 1300 ms, and 1600 ms, meaning the time taken for the collision to occur if the balls had not paused. Additionally, in the two-object condition, one ball moved twice as fast as the other to prevent the use of simple heuristics such as linear extrapolation. We recruited 50 participants online via Prolific, each of whom completed a set of 72 trials (36 trials per condition). Of these, 12 Participants were excluded for failing at least one pre-task qualification question, leaving 38 subjects for analysis.

Results

A generalized linear mixed-effect model with a Gamma distribution for RT showed a significant correlation between the collision time and RT (z-score =7.98, $p < 0.001$). Trials in two-objects condition had longer RTs than the one-object condition, with an average ΔRT of 266 ms (CI = [164, 368]).

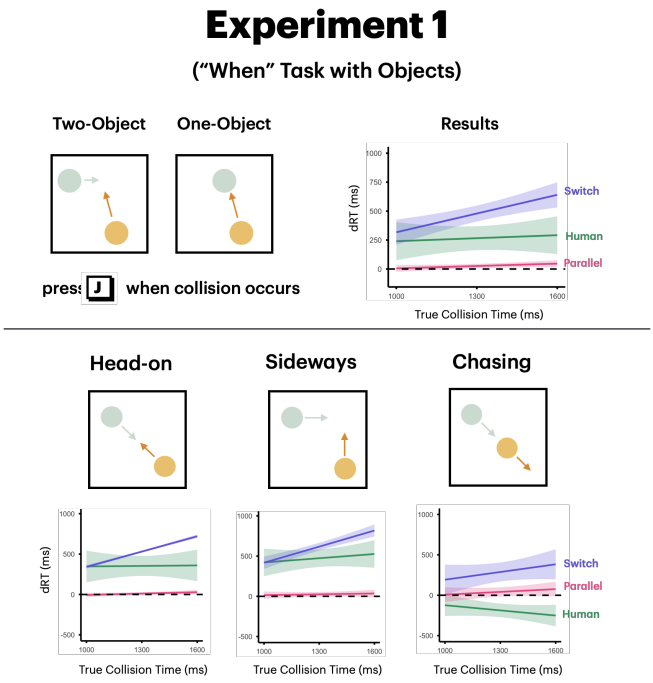

Figure 2: Experiment 1. Top-left: the Two-Objects "When" task. Top-right: Difference in RT between the one-object and two-objects condition (ΔRT) for the human results, the Parallel model and the Switch model. The error bar indicates a 95% Confidence Interval. ΔRT above 0 indicates that simulating two objects was more 'costly' than one object. Bottom: ΔRT comparison between human data and the mental simulation models, split by collision type. The ΔRT for 'chasing' being below 0 may be due to people thinking the collision actually occurred before what the ground truth collision is, see main text and Experiment 2.

[1] We note that Balaban & Ullman (2025) considered a Switch model in their non-interaction stimuli, and found that it did not align with the data. But they did not in principle rule out a Switch model for cases like interactions.

We implemented two mental simulation models to estimate participants' RTs, Parallel and Switch. The Parallel model updates the position of both objects at the same time. The Switch model moves one object for S steps, then the other, back to the first until a collision happens (we considered multiple S values, the results were collapsed across them in the following analysis). In line with the current approach to noisy mental simulation (Battaglia et al., 2013; Balaban & Ullman, 2025), we added noise to the initial position and velocity of the balls to account for perceptual and dynamic uncertainty. We considered various noise levels, but they did not meaningfully alter the model results; thus, we opted to use a noise standard deviation of 0.5 in all the following analyses.

To remove the interference of irrelevant factors like motor delay, we first matched model and human responses in the one-object condition via a linear transformation (that is, people's RT is a linear function of ground truth RT, and this linear function was applied to the models). The Parallel model shows a very small difference in RT between the one-object and the two-objects conditions (average ΔRT = 27.2 ms, CI = [6.38, 45.4]), which is expected, as the Parallel model advances both balls simultaneously, and any difference would be caused by random noise. In contrast, the Switch model predicts a longer RT for the two-objects condition than the one-object condition, with an average ΔRT of 484 ms (CI = [393, 565]). Thus, a Switch model captured the ΔRT in participants' response, while a simple Parallel model failed to do so.

Although participants' RTs were longer in the two-objects condition than in the one-object condition, the change in response time (ΔRT) was much smaller than predicted by the Switch model we considered. To investigate this further, we examined ΔRT across collision types. In the head-on and the sideways collisions, RT in the two-objects condition were longer than in the one-object condition. Conversely, in the chasing collision, RT in the two-object condition was shorter than in the one-object condition. When only considering the head-on and the sideways collisions, ΔRT rises to 417 ms (CI(human) = [339, 488], compared with ΔRT (Parallel) = 18.1 ms, CI(Parallel) = [3.09, 33.1]; ΔRT (Switch) = 575 ms, CI(Switch) = [488, 661]). It is possible that this is because people in the two-object chasing condition believe that the collision happens earlier than it actually does (in such a case, the comparison to the one-object ground-truth is inaccurate). But such an earlier collision would indeed be predicted by a switch model: if the 'chasing' object is the one being simulated while the chased object is maintained statically, it can lead to an early termination of the simulation. We take up this point in the next experiment.

Experiment 2: "Where" Task with Objects

Is the shorter response time in the two-objects condition for the chasing collision type caused by the premature collision predicted by a Switch model? Experiment 2 tested this, by asking participants to report the location of imagined collisions.

Method

The stimuli and procedures were the same as Experiment 1, except for the following: In each trial, instead of pressing a key when the collision happens, participants click *where* the collision would occur. The experiment only contained the two-object condition, as the one-object condition requires a predefined static patch at the collision location, so no spatial prediction is required. We recruited 40 participants online via Prolific, each of whom completed 36 trials. Of these, 9 participants were excluded due to failing at least one pre-specified qualification question.

Results

In head-on and sideways collisions, participants' estimated collision positions showed a roughly Gaussian distribution close to the ground truth, with a slight bias towards the faster ball's approaching direction. However, in the chasing condition, the estimated location shows a much more elongated Gaussian shape, stretching toward the balls' approach direction (see Fig. 3). This indicates that participants thought such collisions occurred much earlier than they actually would have based on the ground-truth parameters.

We considered the same two mental simulation models, but this time to estimate collision positions. The Parallel model predicted collision positions centered on the ground truth location across all collision types, whereas the Switch model predicted the biased estimates observed in the chasing condition. The overlap between such distributions can be formally compared via different methods, generally falling under 'optimal transport' (Peyré et al., 2019). We opted for the simple metric of 'Earth Mover's Distance', akin to moving a pile of earth from one distribution to another. Smaller numbers indicate a better overlap. We used participants themselves to bootstrap the best possible overlap that can be reasonably expected, sampling with replacement and finding a within-participant Earth Mover's Distance of 0.194 (CI = [0.152, 0.236]). In line with Fig.3, the Switch model provided a better fit to participants' data in chasing conditions (Earth Mover's Distance = 0.573, CI = [0.501, 0.646]) than the Parallel model (Earth Mover's Distance = 0.813, CI = [0.646, 0.980]). We did not find a significant difference in the Earth Mover's Distance for different settings of the switch intervals, but it is likely that different people use somewhat different switch intervals, and that we are under-powered to detect such differences between the intervals.

We found that in chasing conditions participants estimated an imagined collision occurring earlier than expected, positioned towards the balls' approaching direction, suggesting an underestimation of the traveled distance in the two-object condition. This result explains the finding in Experiment 1 that RTs were shorter in the two-object condition than in the one-object condition in the chasing collision condition. The intuition underlying this is simple: If one moves the chasing object towards the chased object as the chased object is held

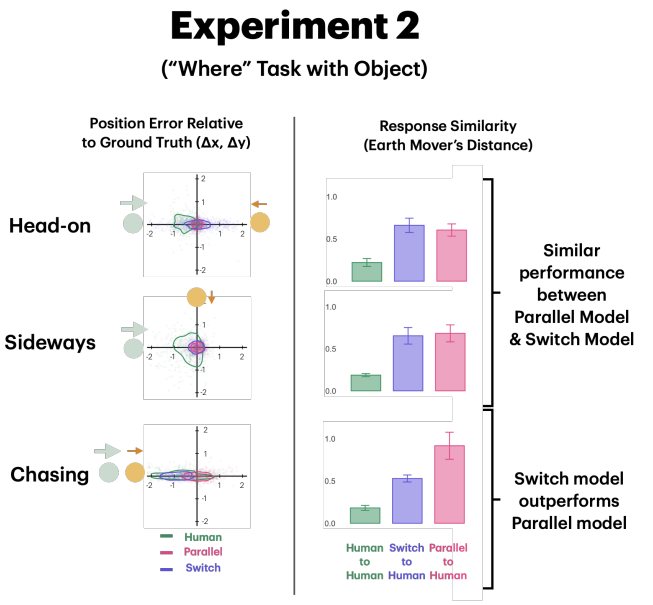

Figure 3: Experiment 2 results. **Left**: Overlaid estimated collision positions from participants, the Parallel model, and the Switch model, broken down by collision types. Two adjustments were applied in order to align the data: (1) each point represents the difference between the estimated collision position and the ground truth collision (the intersection of the two lines), and (2) collision configurations were rotated so that it is as if the faster ball moves from left to right. The units correspond to the original stimulus scale (an 8 × 8 square). **Right**: the Earth-Mover Distance between each model and human data. The error bars indicate the standard error.

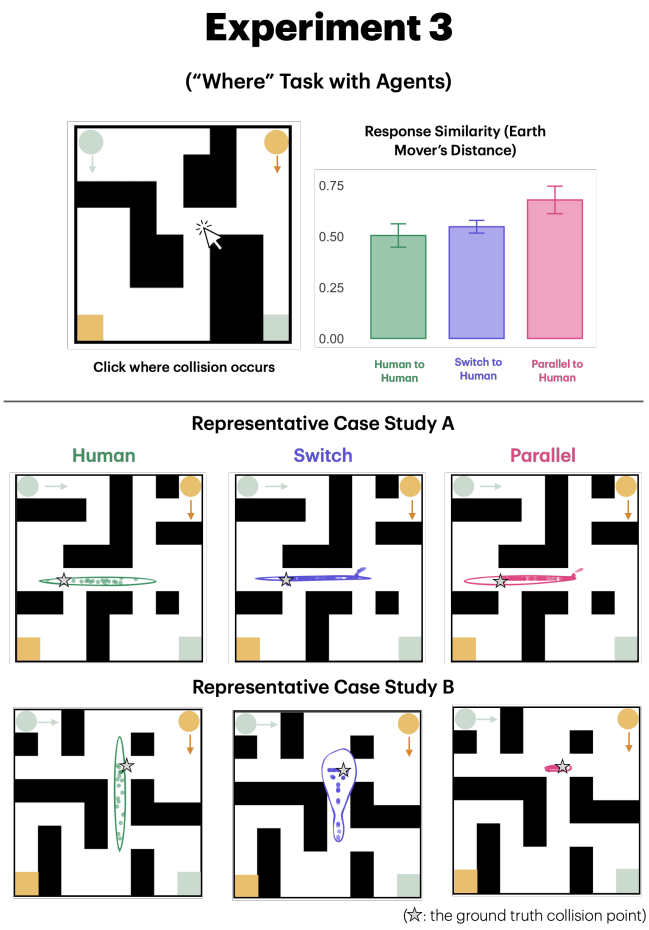

Figure 4: Experiment 3. **Top Left**: Schematic of the task. **Top right**: Results of the experiment. Earth-Mover distance measures similarity between distributions, lower means more similar. The error bars indicate the standard error. The Switch model was closer to Participant data. **Bottom** Example stimuli to highlight model and people behavior. The stars indicate the ground truth collision point.

stationary, it 'catches' it earlier (in space and time) than if the chased ball were moved at the same time as the chaser.

So far, we analyzed the imagined collision of objects in simple linear motion. We next examined the mental simulation of more complicated movements, those of agents. To prevent the bias in collision time caused by incorrect collision position estimation, we first collected people's own judgments of *where* the collision between two agents occurs.

Experiment 3: "Where" Task with Agents

Experiment 1 and 2 considered objects moving along simple paths. In Experiment 3 and 4, we examine the mental simulation of two agents moving along more complex paths on the way to different goals as they navigate a maze. Experiment 3 asked participants to report the estimated collision location for two such agents. This Experiment is the basis for the next experiment as we will use participants' own estimation of the collision as the control, but it also independently helps to distinguish parallel and serial simulation.

Method

In each trial, participants were presented with a 2D maze (see Fig. 4). Each maze depicted two agents (colored circles), two goals objects (squares matching the colors of the agents), and impenetrable walls (black squares). The agents started in separate top corners, and goals were located in contralateral bottom corners. Participants saw agents begin to navigate toward the goals. Both agents disappeared after 1000 ms. Participants were asked to continue the agents' motion in their minds' eye, and click where they believed the agents would collide. For this experiments, we recruited 20 participants online via Prolific. Of these, 1 participant was excluded for failing at least one test question.

We originally created a set of 17 maze stimuli. We note that an agent could in principle take different paths to reach a given goal. Rather than assume a-priori a particular model of planning, we conducted a separate, supplementary experiment (not detailed here). In this supplementary study, we recruited a different group of participants (N=20) who viewed

the maze stimuli, and drew paths an agent would take to reach its given goal. We used these human-based paths as the paths a mental-simulation model should follow, with the only issue being whether these unfold in parallel or not. In doing this, we realized some stimuli allowed agents to take multiple reasonable paths, introducing variability that could not be captured by either of the current models. As these questions of planning are not the focus and do not bear on the main study, we excluded these maze stimuli and analyzed only those with unambiguous paths. This left 10 maze stimuli per participant.

Results

Both participant and the models generated various distributions of collision locations. We can once again consider the question of which model better predicts human as a question of optimal transport, and specifically use Earth Mover Distance to measure which model's output distribution overlaps more with people. We once again used bootstrapping to sample participant data with replacement in order to create a baseline of the best degree of similarity that can be reasonably expected by any model, finding a within-participant Earth Mover's Distance of 0.504 (CI = [0.388, 0.620]). As shown in Fig. 4 (top right), The Switch model provided a better fit to participants' data (Earth Mover's Distance = 0.547, CI = [0.484, 0.610]) compared to the Parallel model (Earth Mover's Distance = 0.678, CI = [0.541, 0.815]). Again, we did not find a significant difference in the Earth Mover's Distance for different settings of the switch intervals.

It is interested to consider in more detail the specific cases in which the Switch model provides a better fit to participants. To see this, consider Fig.4 (bottom), where we highlight two representative examples. In the first case (upper), participants and both models predict the agents will meet in the center of the middle horizontal arm of the maze, with a somewhat 'smeared' distribution within this maze. Note that all these predictions are different from the actual ground truth collision predicted by optimal planning (indicated by the star). In the second case (lower), participants predict that a collision will occur further down from the ground truth collision, along a vertical corridor. The Switch model accounts for this, as there are cases where a given agent will 'miss' the collision point and continue down the corridor, but will be 'caught up' by the second agent when it becomes that agent's turn to be simulated. Unlike these predictions, the Parallel model predicts a distribution centered close to the true collision point (if one object misses the other in parallel simulation, it will not catch up).

Experiment 4: "When" Task with Agents

In our final study, we examined whether the Parallel or the Switch model best predicts participants' estimates of the *timing* of the collision between two moving agents.

Method

The stimuli in Experiment 4 were generally similar to that of Experiment 3. In each trial, participants saw 2D mazes

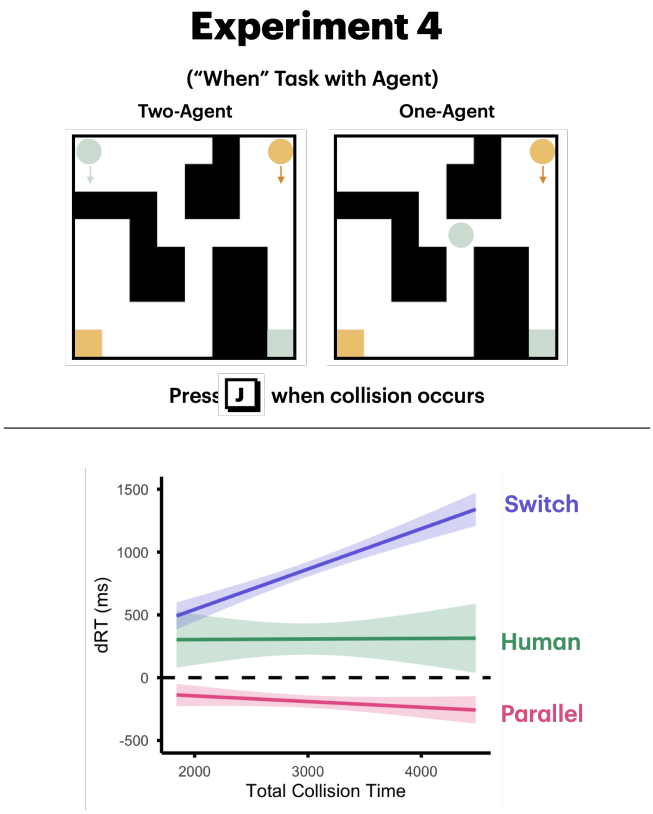

Figure 5: Experiment 4 results. Top: Two-Agents "Where" task. Bottom: difference in RT between the one-agent and two-agents condition for human responses, the Parallel and the Switch models.

in which agents (circles) moved towards goals (colored squares). There were two conditions: In the two-agent condition, two agents started from the top corners and navigated toward separate goals at the opposite bottom corners; both disappeared after 1000 ms. In the one-agent condition, a single agent traveled from a top corner toward a goal at the opposite bottom corner and disappeared after 1000 ms; a static patch was placed at the average collision position found in Experiment 3. This ensured that any RT differences between conditions were not driven by biases in the estimated travel distance. In both conditions, participants pressed a key when they believed a collision occurred, either between the two agents or between the agent and the static patch. We recruited 20 participants online via Prolific. Of this, 6 participants were excluded for failing at least one test question or failing to submit a complete dataset. 14 participants remained.

Results

A generalized linear mixed-effect model with a Gamma distribution for RT revealed a significant correlation between collision time and RT (z-value = 7.14, $p < 0.001$). Trials in the two-agent condition had a longer RT than the one-agent condition, with an average difference of 307 ms (CI = [186,

429]). The Parallel model shows a negative RT difference, with shorter RT in the two-agent condition than in the one-agent condition (ΔRT = -189 ms, CI = [-240, -139]), likely due to differences in estimated collision position between the two conditions. As shown in Experiment 3, the Parallel model exhibits symmetric horizontal error but asymmetric vertical error, with bias only in the upward direction. Because agents start from the top corners and move downward, this pattern suggests that the Parallel model underestimates traveled distance in the two-agent condition, leading to shorter predicted RTs. On the other hand, the Switch model's predicted RT for the two-agent condition was longer than participants (ΔRT = 857 ms, CI = [746, 968]). Such a difference is not symmetric: a positive ΔRT in both the participant and Switch models indicates in favor of the Switch model. However, we do note a more important deviation between the current Switch model and people: As shown in Fig.5, the ΔRT for the Switch model had a positive slope, whereas for people it was relatively flat. We take up this discrepancy in the General Discussion.

General Discussion

When we imagine things, many of us have the experience that we watch something like a movie unfold before our eyes. But research suggests this 'movie' may be more patchwork, and its unfolding more limited than we realize. Such patches and approximations make sense, for a system that is tasked with creating good-enough worlds with limited resources. In particular, previous research suggests a capacity limit in mental simulation, such that people can only update the motion of a single entity, object, or summary statistic at a time Balaban & Ullman (2025), mentally unfolding its trajectory in a serial fashion. But, such a serial simulation presents a problem: How is it that people are able to imagine the interactions between objects?

We investigated how people simulate interactions between multiple objects under a potentially serial constraint. Across four experiments that examined both simple physical motion and goal-directed agent-based motion, we found that a **Serial-Switch model** better accounts for the mental simulation of interactions then a Parallel updating model. In the Switch model, the mind alternates between objects, updating one for some steps before switching to the other, until a stopping condition is met. Consistent with this account, participants' response times increased when they needed to simulate an additional object (Experiments 1 and 4). Moreover, the Switch model uniquely captured systematic biases in estimated collision positions that the Parallel model failed to explain (Experiments 2 and 3). Together, these results suggest that the single-object update-limit in mental simulation holds even when objects can interact, and people simulate these interactions not by moving all objects simultaneously in imagination, but by alternating between them.

Not detailed in our analyses above, in addition to the Parallel and Switch models, we considered a third kind of Serial model, in which the mind first computes the collision point (e.g. via geometric considerations) and then simulates each object fully toward that location. This model is unlikely for two reasons. First, it predicts that collision-point computation precedes simulation, yet RTs in the "where" tasks (Experiments 2 and 3) were comparable to those in the "when" tasks (Experiments 1 and 4). Second, it predicts collision estimates centered on the ground truth, contrary to the systematic biases observed in Experiments 2 and 3. Thus, the full pattern of results favors the Switch model.

Although the Switch model we considered correctly predicts the direction of the response time difference across conditions, the *slopes* of participants' ΔRTs were smaller than predicted. There are different possibilities for why this is the case, and resolving this will require further studies. For example, it is possible that participants applied corrective strategies to compensate for overestimated collision times, such as speeding up simulated motion, or producing faster responses intentionally (as a external corrective on top of simulation). It may also be that the reduced difference is due to accumulated noise during longer simulations. For example, in Experiment 4, the standard deviation of responses was significantly greater for longer ground-truth collision times, which may have masked differences between conditions.

The Switch model we considered made several choices that can be investigated further. In particular, we averaged over different switch intervals. However, it seems possible that different participants use different intervals, and also adjust them in a context-based way. Though not explored in the current study, switch intervals may even vary within a trial. For example, people may alternate between objects more rapidly the closer an anticipated collision gets. Future work could examine how switch intervals are selected, how they adapt to task demands, and whether they dynamically change as a simulation unfolds.

Our work extended the previous examination of non-interacting physical objects to both simple physical interactions and simple agent-based motions. However, further work should move beyond simple objects and agents to more complex, real world scenes and more complex, real world social interactions and goals. The extension to complex social scenes is not meant just as a question of agent motion, but also the process of updating internal agent states. As just one example, when viewing a scene in which one agents helps or hinder another and simulating their future behavior, how do observers simulate the updates to agent beliefs and intentions? Is this done in parallel, or serial, modeling the change to one mental state at a time?

Imagine a rugby player tackling another running player. Two seals do a high five. The rotating arm of a windmill swats a passing witch off her broom. We can imagine a variety of familiar and fantastical interactions. But while our imagination can run free and wild, it cannot run two things in tandem. Rather, it ferries back and forth, fast and fleeting; switching ends in objects meeting.

References

Alvarez, G. A., & Franconeri, S. L. (2007). How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism. *Journal of Vision*, *7*(13), 14.

Balaban, H., & Ullman, T. (2025). The capacity limits of moving objects in the imagination. *Nature Communications*.

Bass, I., Smith, K. A., Bonawitz, E., & Ullman, T. D. (2021). Partial mental simulation explains fallacies in physical reasoning. *Cognitive Neuropsychology*, *38*(7-8), 413–424.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*, 61–76.

Lovett, A., Bridewell, W., & Bello, P. (2019). Selection enables enhancement: An integrated model of object tracking. *Journal of Vision*, *19*(14), 23.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, *11*(5-6), 355–607.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, *3*(3), 179–197.

Smith, K., Battaglia, P., & Tenenbaum, J. (2023, July). *Integrating heuristic and simulation-based reasoning in intuitive physics* (preprint). PsyArXiv. Retrieved 2024-03-04, from `https://osf.io/bckes` doi: 10.31234/osf.io/bckes

Sosa, F. A., Gershman, S. J., & Ullman, T. D. (2025). Blending simulation and abstraction for physical reasoning. *Cognition*, *254*, 105995.

Spelke, E. (2022). *What babies know: Core knowledge and composition volume 1* (Vol. 1). Oxford University Press.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind Games: Game Engines as an Architecture for Intuitive Physics. *Trends in Cognitive Sciences*, *21*(9), 649–665.

Wang, Y., & Ullman, T. D. (2025). Resource bounds on mental simulations: Evidence from a liquid-reasoning task. *Journal of Experimental Psychology: General*.