Seeing through Occlusion: Uncertainty-aware Joint Physical Tracking and Prediction

Arijit Dasgupta[1], Andrew D. Bolton[2], Vikash K. Mansinghka[1], Joshua B. Tenenbaum[1], Kevin A. Smith[1]
[1] Massachusetts Institute of Technology, Cambridge, Massachusetts, United States
[2] CHI FRO, Cambridge, Massachusetts, United States
Correspondence to arijitdg@mit.edu

Abstract

Humans can track objects and predict their motion even when they are temporarily occluded. How does the absence of changing visual evidence alter predictive beliefs about a moving object? In our study, participants were tasked with continuously anticipating the destination of a simulated ball in occluded and un-occluded 2.5D environments. Our findings reveal that humans actively update their judgments throughout the period of occlusion while making predictions grounded in physical realism, even as occlusion impairs accuracy. To model this behavior, we integrate perception with physical reasoning, unifying tracking and prediction. This is implemented via massively parallel probabilistic inference in a hierarchical generative model for the motion of intermittently visible objects, represented using the GenJAX probabilistic programming platform. This model predicts time-varying human judgments more accurately than alternative models, suggesting that humans integrate perception and physics to reason about occluded motion.

Keywords: Physical Reasoning; Probabilistic Reasoning; Tracking and Prediction; Occlusion

Introduction

Imagine racing across the rink, eyes locked on the puck, before it vanishes behind a cluster of ice hockey players, blocked from view. The instinct is not to stop or hesitate, but to keep moving, anticipating where the puck will reappear—to go "where the puck is going to be, not where it has been," as Wayne Gretzky wisely put it. This is more than just practical sporting advice; it mirrors the way we navigate the world. Sensory information helps us understand what is present and how it changes, yet humans can maintain internal representations of occluded motion and continue to make inferences despite missing visual input. In the face of occlusion, our minds don't simply pause—they rely on internal models, drawing from past observations, beliefs, and physical laws to predict what will come next.

The ability to reason under occlusion is a fundamental capacity that emerges early in human development. Even infants (Kellman & Spelke, 1983; Baillargeon, 1987) represent static, occluded objects as coherent and permanent. Humans can also maintain internal representations of occluded moving objects (Scholl & Pylyshyn, 1999), directing greater attentional resources toward them (Flombaum et al., 2008). Furthermore, studies on prediction motion (Tresilian, 1995) reveal that humans *actively* track a single object during periods of occlusion (DeLucia & Liddell, 1998; Makin & Poliakoff, 2011). Yet, these investigations often focus on brief occlusions (< 1s) and simple, unobstructed motion, leaving open the question of how humans reason about longer occlusion periods, or settings where occluded objects may collide.

Occlusion also introduces an inherent uncertainty in human judgment of motion. Research by Lyon and Waag (1995) and Makin and Chauhan (2014) on production tasks demonstrates that as occlusion duration increases, human performance in predicting when an object will reappear diminishes, with errors in timing estimates growing linearly as occlusion persists. This suggests that uncertainty not only arises with occlusion, but also deepens as time progresses. ***How, then, does this uncertainty influence people's judgments of occluded motion when potential collisions are involved?***

A central modeling assumption lies in the integration of physical dynamics with probabilistic reasoning. The "intuitive physics engine" framework by Battaglia et al. (2013) offers a conceptualization of the physical world as a structured probabilistic generative process, showing that forward simulations of noisy inputs can capture human judgments across various tasks involving physical interactions, such as predicting the stability of a stacked tower of blocks. Similarly, Sanborn et al. (2013) showed that optimal Bayesian inference over a statistical collision model explained human judgments of mass-collision events. For time-varying predictions, Smith et al. (2013) demonstrated how noisy forward simulations of physical dynamics aligns with human judgments of future object destinations. Their approach, though effective with full state information, is limited in occluded situations as they rely on ground-truth data and lack an observation likelihood model. Further work by Smith et al. (2019) addressed expectation violations in intuitive physics, modeling human-like surprise at non-physical events like object disappearances. While their model represented physical latents like position and velocity, it focused primarily on inferring object existence.

Much of prior work investigates tracking and prediction as separate problems. In contrast, scenarios that involve reasoning through occlusion—such as following a moving object behind a barrier or predicting the trajectory of a ball when part of its path is hidden—require both: keeping track of where an object *is* and predicting where it *will be*. We propose that these processes must be integrated for effective reasoning, as uncertainty in one process naturally propagates to the other.

To address these gaps, we propose **J**oint **T**racking **A**nd **P**rediction (**JTAP**), a computational model that combines perception, tracking and prediction as a unified approach to men-

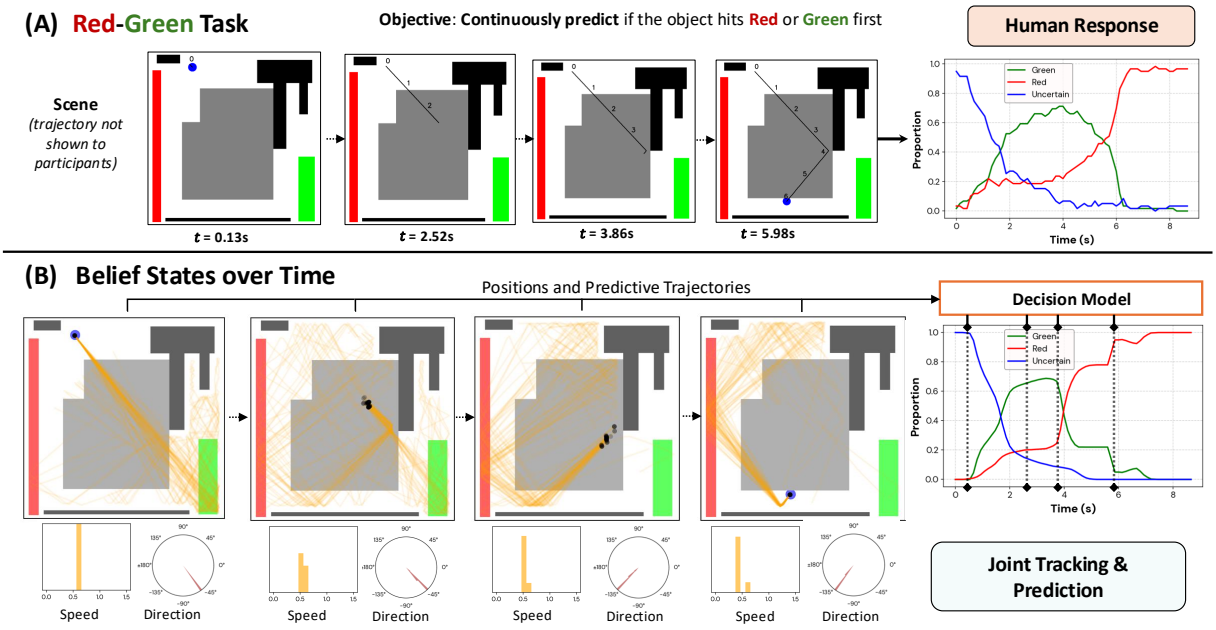

Figure 1: (A) Participants were asked to continuously judge if a moving ball (**Blue**) will hit Red or Green first in a 2.5D environment with occluders (**Gray**) to hide ball motion and Barriers (**Black**) for possible collisions. (B) Time-varying interpretable belief states from one run of **JTAP**. The black dots depict the ball's positional posterior estimate while the yellow lines illustrate the ball's predictive posterior over future trajectories. Multiple runs of **JTAP** are used by a decision model to make varying individual-level predictions of red and green, which are then aggregated and compared with participants.

tal physical reasoning. By extending the 2.5D bumper table domain from Smith et al. (2013) to include occluders, our model bridges the gap between tracking an unseen object and predicting it's future state, combining probabilistic inference with physically realistic simulations. Our approach maintains interpretable probabilistic beliefs about an object's position, speed, and direction at every timestep of its motion (Figure 1(B)), and uses this information to make discrete predictions about future outcomes. **JTAP** thus provides time-varying **individual-level** predictions of an object's future destination.

To test **JTAP**, we conducted a behavioral experiment requiring participants to judge the future destination of a moving ball (Figure 1(A)). By varying scene properties, like the existence of occlusion, object motion and placement of barriers, we found that human predictions continue to be guided by expectations about the current location of the ball, even as they become less accurate under occlusion. By aggregating multiple individual-level predictions, **JTAP** predicts the same reasoning patterns, outperforming plausible alternative hypotheses. These findings suggest that, like human predictions, **JTAP** adapts its judgments to align with physically plausible predictions, reflecting the graded uncertainty that arises in the absence of direct visual evidence of the moving ball.

The Joint Tracking and Prediction Model

We designed **JTAP** as a probabilistic modeling and inference method to test the hypothesis that people jointly track and predict the motion of an object in a 2.5D environment, where occlusions may occur, based on the same video observations available to the observer (Figure 1). At its core, **JTAP** assumes people maintain a time-varying representation of the world state, integrating perception and physical dynamics as a means of intuitive physical reasoning (Smith et al., 2024). The state at any time step is described by the object's position, speed, and direction, with state evolution governed via a stochastic structured dynamics model that incorporates uncertainty about motion and collisions.

To reason about these states, and similar to previous probabilistic object tracking work (Vul et al., 2009; Farahi & Yazdi, 2020), **JTAP** uses Bayesian inference to infer the world state X_t given all observations $Y_{0:t}$, but extends this framework to **jointly** predict possible future states $X_{t\ldots t+M}$. With each new observation, **JTAP** jointly and approximately updates its current belief state and predictive future states of the object via an observation likelihood model that upweights prior beliefs that best explain new observations. During the visible motion of the object, **JTAP** infers the object's state and uses that internal representation to continue tracking the object when it becomes hidden, accounting for the uncertainty introduced by the absence of visual evidence. Furthermore, **JTAP** is capable of relying on its internal dynamics and inferred knowledge to predict possible points of collisions with barriers placed at the edge of occluders, even when these collisions cannot be seen.

Probabilistic Model

Our model is assumed to operate in a known and static 2.5D environment, with the positions and sizes of barriers and occluders extracted deterministically from RGB videos of the scene. The state of the object, X_t, at time t is defined by the x and y positions (S_{x_t}, S_{y_t}), speed (v_t) & direction (ϕ_t) of the object. Observations are represented by a discrete-valued image Y_t, where each pixel $\Psi_{ij} \in \{0, \ldots, K-1\}$ corresponds to one of K possible values (i.e. empty space, object, barrier, occluder) in the ij position. The latent state and observation evolves as per the following generative process.

$$\textit{Initial State Prior: } X_0 \sim P(X_0) \tag{1}$$

$$\textit{Observation Model: } Y_t \sim P(Y_t|X_t, \rho) \tag{2}$$

$$\textit{Physical Dynamics Model: } X_t \sim P(X_t|X_{t-1}, \eta) \tag{3}$$

Initial State Prior The initial state, X_0 is sampled from a broad uniform prior over all possible states. The object's position is sampled within the scene dimensions, its direction from any angle, and its speed up to a maximum speed.

Observation Model Observations are first generated using a deterministic rendering function from the object state, $\hat{Y}_t = \textit{Render}(X_t)$, and then each pixel is independently sampled with a ρ probability of being corrupted to one of $K-1$ other colors. As we constrain on images during inference, the observation model behaves as a likelihood evaluator: $P(Y_t = y_t|X_t = x_t, \rho) = \mathcal{L}(x_t, \rho; y_t) = \prod_{i,j}\left[(1-\rho)\mathbb{I}[\psi_{ij} = \hat{\psi}_{ij}] + \frac{\rho}{K-1}\mathbb{I}[\psi_{ij} \neq \hat{\psi}_{ij}]\right]$. Here, ψ_{ij} refers to the observed pixel at position (i, j), and $\hat{\psi}_{ij}$ is the corresponding pixel from the rendered image $\hat{Y}_t$.

Physical Dynamics Model The physical dynamics model incorporates a collision-aware model of 2D motion, with noise in direction, speed, and position as shown in Figure 2. The noise model is inspired by the work of Smith and Vul (2013) and is used to approximate the logical uncertainty of computing future physical states. We assume dynamic noise $\eta = \{\sigma_{Col}, \sigma_{Dir}, \Sigma_S, \sigma_\mathrm{v}\}$, including noise in how collisions resolve (σ_{Col}), noise in the direction the ball travels (σ_{Dir}), positional uncertainty (Σ_S), and uncertainty in the speed (σ_v). This generates samples of expected future world states from prior states and noise: $P(X_t|X_{t-1}, \eta)$.

Joint Tracking and Prediction

To perform joint tracking and prediction, we first define tracking as a Bayesian filtering problem (Särkkä & Svensson, 2023), where the posterior distribution of interest is the filtering posterior, $P(X_t|Y_{0:t}, \rho, \eta)$. Prediction is defined by the predictive posterior, $P(X_{t+M}|Y_{0:t}, \rho, \eta)$, which extends the filtering posterior M timesteps into the future. We implement a Sequential Monte Carlo algorithm (Del Moral et al., 2006) as described in Algorithm 1. At each timestep t, the algorithm outputs a set of N weighted particles, $\{(w_t^i, x_{t:t+M}^i)\}_{i=1}^N$, representing the belief state of the current state and the future trajectory M timesteps ahead. The filtering and predictive posteriors are discretely approximated as:

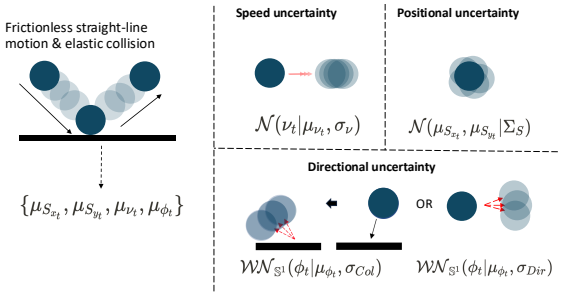

Figure 2: **The physical dynamics model**: mean values for speed (μ_{v_t}), position $(\mu_{S_{x_t}}, \mu_{S_{y_t}})$, and direction (μ_{ϕ_t}) are derived from frictionless straight-line motion and elastic collision. Speed uncertainty is modeled as a Gaussian distribution $\mathcal{N}(\mathrm{v}_t|\mu_{\mathrm{v}_t}, \sigma_\mathrm{v})$, positional uncertainty as a 2D Gaussian $\mathcal{N}(\mu_{S_{x_t}}, \mu_{S_{y_t}}|\Sigma_S)$, and directional uncertainty as a Circular Wrapped Gaussian $\mathcal{WN}_{\mathbb{S}^1}(\phi_t|\mu_{\phi_t}, \sigma_{\mathrm{Col/Dir}})$, depending on whether an expected collision occurs.

$$P(X_t|Y_{0:t}, \rho, \eta) \approx \sum_{i=1}^{N} w_t^i \cdot \delta(x_t - x_t^i) \tag{4}$$

$$P(X_{t+j}|Y_{0:t}, \rho, \eta) \approx \sum_{i=1}^{N} w_t^i \cdot \delta(x_{t+j} - x_{t+j}^i) \text{ for } j = 1, \ldots, M \tag{5}$$

As described in Algorithm 1, the proposal distribution, $Q(X_t|X_{0:t-1}, Y_{0:t})$, plays an important role in updating the model's beliefs at each timestep by generating candidate states based on past observations and states. Because each particle represents a discrete hypothesis about the world state X_t, the proposal distribution allows **JTAP** to sample a single plausible world state, focusing this sample on values that are more likely to be plausible given the prior state and current observation, emulating the way the human mind directs attention to the most recent evidence while disregarding less likely possibilities. As illustrated in Figure 3, **JTAP** constructs the proposal distribution using a combination of enumerative grid inference and data-driven estimation. For positions, we discretize the scene into a uniform 2D grid, where each cell corresponds to a fixed spatial area. All grid positions are evaluated under the probabilistic model to compute a grid of log probabilities, which is then used to sample a cell via a categorical distribution. The final position is drawn uniformly from within the sampled cell's spatial bounds.

To propose speed and direction, **JTAP** extrapolates the currently inferred trajectory by identifying the last visible segment before occlusion or the last point of collision. It then linearly projects forward the expected speed and direction from this segment. These projected values define the mean parameters for a wrapped Gaussian (for direction) and a Gaussian (for speed), from which new samples are drawn. During occlusion, the observation model provides no informative

Algorithm 1 Joint Tracking and Prediction

Input: Observations $y_{0:T}$, noise parameters η, ρ, particle count N, resampling threshold $N_{\text{threshold}}$, prediction steps M.
Output: Weighted particles $\{w_t^i, x_{t:t+M}^i\}_{i=1}^N$ for $t = 0, \ldots, T$.

Initialize particles $\{x_0^i\}_{i=1}^N$ by proposing $x_0^i \sim Q(X_0|y_0)$
$\tilde{w}_0^i = \frac{\mathcal{L}(x_0^i, \rho; y_0) \cdot P(x_0^i)}{Q(x_0^i|y_0)}$, $i = 1, \ldots, N$
Self-Normalize: $w_0^i = \tilde{w}_0^i / \sum_{j=1}^N \tilde{w}_0^j$, $i = 1, \ldots, N$
for $t = 1$ to T **do**
▷ Tracking Phase
  for $i = 1$ to N **do**
    Propose $x_t^i \sim Q(X_t|x_{0:t-1}^i, y_{0:t})$
    $\tilde{w}_t^i = w_{t-1}^i \cdot \frac{\mathcal{L}(x_t^i, \rho; y_t) \cdot P(x_t^i|x_{t-1}^i, \eta)}{Q(x_t^i|x_{0:t-1}^i, y_{0:t})}$
  end for
  Self-Normalize: $w_t^i = \tilde{w}_t^i / \sum_{j=1}^N \tilde{w}_t^j$, $i = 1, \ldots, N$
  $N_{\text{eff}} = 1 / \sum_{i=1}^N (w_t^i)^2$
  if $N_{\text{eff}} < N_{\text{threshold}}$ **then**
    for $i = 1$ to N **do**
      Resample $j \sim \text{Categorical}(\{w_t^j\}_{j=1}^N)$
      $x_t^i \leftarrow x_t^j$, $w_t^i = \frac{1}{N}$
    end for
  end if
▷ Prediction Phase
  for $j = 1$ to M **do**
    for $i = 1$ to N **do**
      Sample $x_{t+j}^i \sim P(X_{t+j}|x_{t+j-1}^i, \eta)$
    end for
  end for
  Save weighted particle set: $\{(w_t^i, x_{t:t+M}^i)\}_{i=1}^N$
end for

evidence, so the proposal defaults to the physical dynamics model governed by the simulation noise parameters. This yields samples consistent with the physical dynamics model but with appropriate uncertainty, enabling **JTAP** to maintain plausible roll-outs despite the absence of direct visual input.

While **JTAP** is inspired by the forward simulation approach of Smith et al. (2013), it differs fundamentally in purpose and structure. Smith et al. (2013)'s model assumes access to the ground-truth object state at every timestep and performs noisy forward rollouts from these known inputs. As a result, it does not perform inference, lacks an observation model, and cannot reason under occlusion. In contrast, **JTAP** performs joint inference over both current and future latent states directly from noisy visual input, using a principled observation model to update beliefs even when the object is hidden.

The Red-Green Task

To evaluate **JTAP**, we conducted a behavioral experiment designed to test human tracking and prediction.

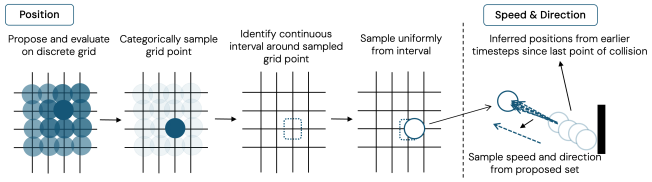

Figure 3: The probabilistic proposal implemented in **JTAP**. At each time step, each particle must update its discrete representation of the variables representing the state of the world. Positions (*left*) are updated by forming a grid around the prior position, then evaluating the probability of the ball being at each grid position. One grid position is sampled relative to these probabilities, and that position is uniformly perturbed within a small continuous interval to produce the proposal. The speed and direction (*right*) are sampled around the speed and direction imputed from the last observed positions from the point of last inferred collision.

Behavioral Experiment

Procedure Participants completed trials where a frictionless simulated ball with elastic collision dynamics moved in a 2D environment with barriers and occluders. Each trial ended when the ball reached a red or green region. Participants continuously predicted the ball's destination, with a score based on accuracy and speed, following Smith et al. (2013): $Score = 20 + 100 * (Prop(Correct) - Prop(Wrong))$. This encouraged quick, accurate decisions while penalizing wrong choices when uncertain. Participants were shown the scene only from the start of each trial, with no prior context and they predicted the ball's final destination by pressing either 'F' for **Red** or 'J' for **Green**. No button press or both buttons pressed were treated as no decision.

Stimuli A total of **50** trials were generated as stimuli, presented as 400x400 pixel videos at 30 FPS with ball motion simulated using `PyMunk` (Blomqvist, 2024). Each video featured black rectangular barriers, gray occluders partially or fully blocking visual access for varying durations ($T_{\text{occlusion}}$), and rectangular red and green goal regions with a blue circular ball (10px radius). Of the trials, 34 included occlusion, and 16 had un-occluded paths. We included four catch trials with obvious outcomes (e.g. the ball was always moving directly at one goal and away from the other). Trial durations ranged from 4.17s to 9.17s, with $T_{\text{occlusion}}$ between 2.90s and 8.13s. To counterbalance color bias, the goal colors were randomly shuffled. Data were corrected during post-processing, and trial order was randomized for each participant.

Participants We recruited 60 participants (mean age 37.69; 35 male, 25 female) via Prolific. All participants had normal or corrected-to-normal vision, and no colorblindness. They were compensated US$15/hour, and the experiment took approximately 14 minutes. Participants completed three familiarization trials before the main experiment. One participant's

data was excluded for scoring below 40 on 3 of 4 catch trials.

JTAP Implementation

To implement a model of the Red-Green task, we assume that **JTAP** represents the evolving beliefs of an individual person about the current and future state of the ball's motion. However, as described in the next section, people must make a discrete decision at each time point about whether and which goal to indicate that the ball will reach. Therefore we apply a deterministic decision model that uses the raw belief states to get individual pseudo-participant key press decisions, and aggregate those decisions over multiple model runs to match to aggregate human data.

Raw Beliefs We estimate the red-green outcome, $P(\zeta_t)$, where $\zeta_t \in \{\text{Red}, \text{Green}, \text{Uncertain}\}$, by checking the predictive outcomes of all particles and using the weighted predictions to approximate this raw belief at each timestep.

Decision Model We implement a decision model by assuming that **JTAP** will make a discrete prediction that the ball will hit the red or green goal when the raw belief about that future event exceeds a threshold, θ_{press}. Since people cannot immediately process a scene and issue motor commands to press a button, we align model and human decisions by introducing a delay, τ_{delay}, in **JTAP**'s decision-making. To capture decision stickiness (as people are unlikely to change their decision based on momentary fluctuations in belief), we define minimum time periods, τ_{press} and τ_{release}, where a sustained belief above the threshold triggers a press, and a sustained belief below the threshold triggers a release. To model the decision making variability between humans, these parameters are sampled from a Gaussian distribution. This mechanism also captures two key forms of uncertainty that lead to people pushing the "Uncertain" button. Ignorance arises when predictive particles do not reach either goal—causing all beliefs to remain below threshold—leading the model to stay in the "Uncertain" state. Equivocation arises when predictive particles are split between outcomes, keeping both red and green beliefs near threshold but not dominant, also resulting in no keypress. These uncertainty types are resolved through the combined effects of thresholding, press/release timing constraints, and delayed response.

Alternative Hypotheses To evaluate the hypothesis that joint tracking and prediction *during occlusion* is a crucial human-like process, we introduced two alternative hypotheses to **JTAP** as ablative baselines. The *frozen model* keeps red-green beliefs ($P(\zeta_t)$) constant during the period of occlusion, testing the hypothesis that people stop tracking and hold their beliefs until the object reappears. We define the period of occlusion as any stimulus frame the object is not fully visible in. The *decaying model* exponentially decays these belief states to *Uncertain* with a decay constant of $2.67s$ in that same time period, testing if humans become more uncertain about the object's destination as occlusion duration increases.

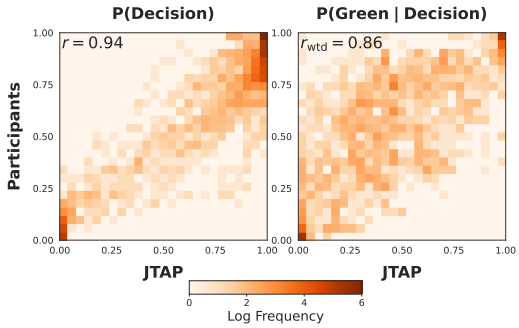

Figure 4: Joint histogram of **JTAP** (x-axis) and human (y-axis) decisions across all trials for each timestep. *Left:* probability of making a decision, either red or green. *Right:* probability of choosing green given a decision was made. Colors indicate the log-frequency of time points within each bucket; redder means more observations. Data along the diagonal indicate good match between human and model predictions.

Computational Details All modeling and inference are implemented as a sequential probabilistic program using `GenJAX` (Becker et al., 2024), a GPU-accelerated probabilistic programming framework. Observations are sampled in intervals of 133ms (1 in every 4 frames), reducing computational cost while capturing relevant object motion. For each trial, 100 runs of **JTAP** and the alternates are generated with different random seeds to obtain 100 sets of raw beliefs. Noise parameters are tuned individually for each computational approach—**JTAP**, the frozen, & decaying models—via grid search, optimizing for the Root Mean Squared Error (RMSE) against red, green, and uncertain levels across all timesteps from participant data. A particle count of $N = 25$ was the best fit for all 3 models. Similarly, the mean and standard deviation for the Gaussian distributions of the decision model parameters were tuned via grid search. To approximate expected decisions, 100 decision model parameter sets are sampled per run of **JTAP** or alternates, resulting in 10,000 pseudo-participants, with the same parameters used across all trials.

Results

Human Performance As the difficulty of each trial varied significantly, the participants' mean score varied from 11.3 to 97.9. Nonetheless, participants' predictions were reliable across trials (Intraclass Correlation = 0.952; Shrout and Fleiss (1979)). Participants mean performance was significantly better ($t(48) = 3.94$, $p < 0.001$) in trials without occlusion (Mean = 64.5) than trials with occlusion (Mean = 38.9).

Model Performance To analyze the performance of **JTAP**, we examine both the probability of making a decision by pressing either the green or red button: **decision** ($P(\textit{Red or Green})_{ij}$) and the conditional probability of choosing green, given a decision to press has been made: **choice** ($P(\textit{Green}|\textit{Red or Green})_{ij}$), across all timesteps (i) in all trials (j). To account for the reliability of participant data at

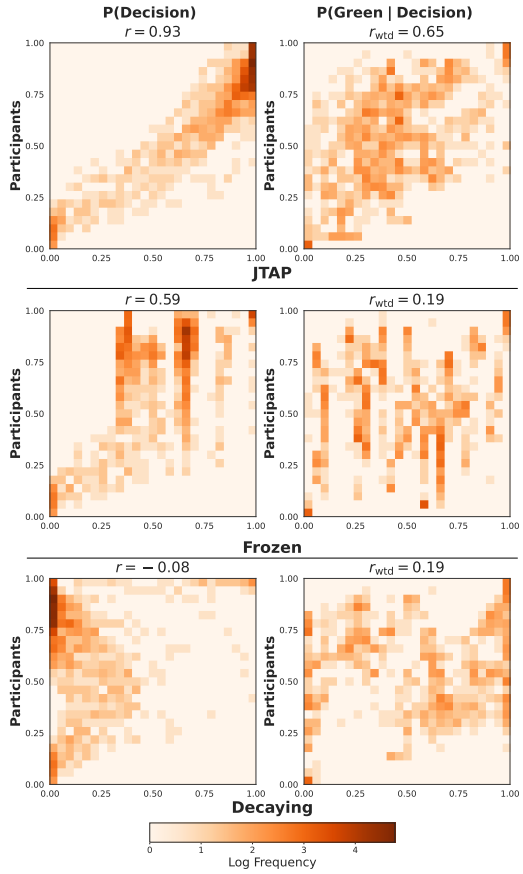

Figure 5: Joint histogram of model and human decision for **occluded** time points only, split by **JTAP** model and alternates. During these diagnostic time points, the **JTAP** model explains when people decide to make predictions and which predictions people make better than the alternatives.

any time-point ij, we weight all **pairs** of participant-model **choices** by the proportion of humans who pressed a button. The keypress outputs from **JTAP** showed substantial correlation in both **decisions** ($r = 0.94$) and **choices** ($r_{wtd} = 0.86$). As both **JTAP** and participants are largely unanimous in their choices across many time-steps, we plot the log-frequency of these choices in Figure 4 to focus on time-steps where there is less agreement between participants. The heatmap on the left shows that **JTAP** effectively captures the uncertainty gradations present in participant **decisions**. The heatmap on the right depicts that **JTAP** also exhibits human-like uncertainty in **choice**. Nevertheless, the gradations are more spread out, indicating that while **JTAP** captures the direction of changing beliefs among participants, it occasionally deviates from perfectly time-aligning with the exact probability levels.

Alternative Hypotheses Although we fit model parameters to all trials, we focus on timesteps with occlusion for model comparisons, as we expect **JTAP** and alternatives to perform similarly when the ball is visible. We find that **JTAP** outperforms both alternative hypotheses. The correlation for **decisions** with human keypresses was higher in **JTAP** ($r = 0.93$) compared to the frozen ($r = 0.59, p < 0.001$) and decaying ($r = -0.08, p < 0.001$) models. The left column of heatmaps in Figure 5 reveals that unlike the frozen and decaying models, **JTAP** is capable of matching human decisions when a larger proportion of participants decide to press a button. This effect is worst in the decaying model, where decaying beliefs make it less likely for the model to press a button even when most participants do. The right column of Figure 5 shows that **JTAP** correlates better ($r_{wtd} = 0.65$) than the frozen ($r_{wtd} = 0.19, p < 0.001$) and decaying ($r_{wtd} = 0.19, p < 0.001$) models with humans for **choices**. In contrast, the gradations in both the frozen and decaying models don't match human patterns well, over and under predicting the probability of choosing green frequently. These findings suggest that humans actively reason when the object is occluded. Moreover, since the decision model accounts for reaction time delay, this distinction is better explained by the changing belief states exposed by **JTAP** during occlusion, compared to alternates.

The examples in Figure 6 demonstrate configurations inducing varying reasoning patterns between **JTAP** and the alternatives. Trial A shows a case where the occluder's placement is inconsequential, and people consistently expect the ball to hit green, captured by **JTAP** and the frozen model, but not the decaying model. Trial B exemplifies how different **JTAP** runs can reinforce different beliefs about the outcome under occlusion, with the brief period of occlusion inducing a subtle but similar change in participants and **JTAP** alike. Trial C exhibits a similar reasoning pattern, where participant decisions are well explained by **JTAP** during the period of occlusion, with both red and green beliefs rising. Trial D presents a case where humans gradually predict red after the ball fails to appear on the left side within a certain time period. This reasoning cannot be captured by the alternatives for they do not reason during occlusion. In contrast, **JTAP** maintains noisy yet physically realistic beliefs and predictions throughout occlusion, allowing it to align well with the trend in human data.

Discussion

This research quantitatively characterizes humans' ability to continually update predictions about an object's motion even while it is occluded. **JTAP**, integrating perception, probabilistic reasoning, and physical knowledge, captures time-varying patterns of human behavior. While our work captures human behavior in a simplified 2.5D setting, it serves as a step toward extending these models to 3D reasoning contexts. Future extensions of **JTAP** could incorporate full 3D object geometry and occlusion, enabling applications to more complex physical tasks such as tabletop interaction or robotic planning. Meeting this challenge will require handling high-dimensional state spaces, richer visual input, and more structured uncertainty.

Future refinements could improve alignment with human response timing and variability. In particular, we observe that **JTAP** captures the trend of changing beliefs over outcomes

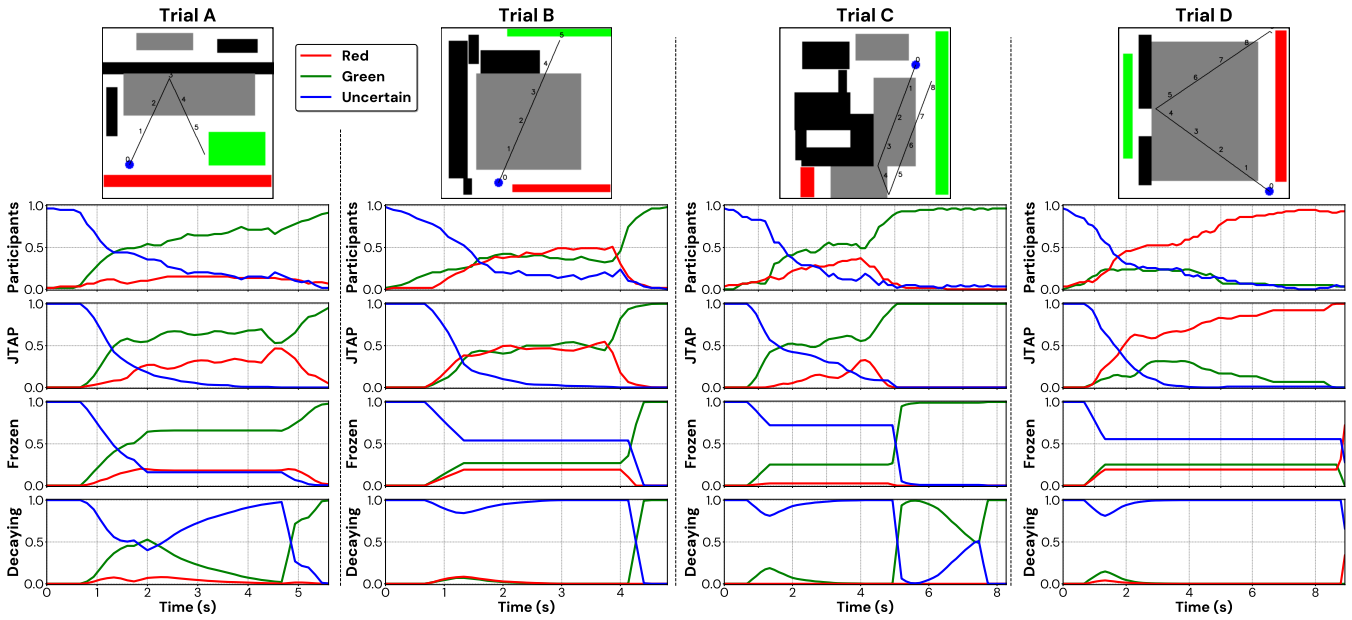

Figure 6: Illustrative examples of models and alternates in comparison with human predictions over four trials (A - D). Trial A shows the limitation of the decaying model on a simple scene while trials B to D demonstrate how joint tracking and prediction under occlusion provide a more compelling account for the subtle yet telling time-varying human responses.

among humans but sometimes struggles to align exactly with the timing or probability levels. This suggests that while our model captures human-like reasoning, incorporating human prior over features like motion direction or assumptions about experimenter demands, could improve its precision. These results show that halting or decaying belief states during occlusion fails to capture hidden collisions and the associated physical reasoning.

Future work could apply this Bayesian framework to other aspects of physical reasoning, such as predicting outcomes when objects are unseen—e.g., if a ball rolls under a couch and doesn't come out, is something blocking it? Additional behavioral experiments could also systematically manipulate scene geometry—such as varying the size, placement, and alignment of barriers, occluders, and goal regions—to analyze how these spatial changes affect human judgment under occlusion, and whether the model captures these fine-grained patterns. And this framework could be extended to support more general tracking and prediction – for instance predicting where a pedestrian walking under and underpass will emerge by using simulations of agent motion (Shu et al., 2021) instead of physical dynamics. These extensions would help test the generality of our approach and reveal which components of intuitive physics are most sensitive to visual uncertainty and scene structure complexity.

Acknowledgments

This work was supported in part by CoCoSys, one of seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA, and by NSF grant 2121009.

References

Baillargeon, R. (1987). Young infants' reasoning about the physical and spatial properties of a hidden object. *Cognitive Development*, *2*(3), 179–200.

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.

Becker, M. R., Lew, A. K., Wang, X., Ghavami, M., Huot, M., Rinard, M. C., & Mansinghka, V. K. (2024). Probabilistic programming with programmable variational inference. *Proceedings of the ACM on Programming Languages*, *8*(PLDI), 2123–2147.

Blomqvist, V. (2024, October 13). *Pymunk*. Retrieved from `https://pymunk.org` (An easy-to-use pythonic rigid body 2D physics library.)

Del Moral, P., Doucet, A., & Jasra, A. (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *68*(3), 411–436.

DeLucia, P. R., & Liddell, G. W. (1998). Cognitive motion extrapolation and cognitive clocking in prediction motion tasks. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 901.

Farahi, F., & Yazdi, H. S. (2020). Probabilistic kalman filter for moving object tracking. *Signal Processing: Image Communication*, *82*, 115751.

Flombaum, J. I., Scholl, B. J., & Pylyshyn, Z. W. (2008). Attentional resources in visual tracking through occlusion: The high-beams effect. *Cognition*, *107*(3), 904–931.

Kellman, P. J., & Spelke, E. S. (1983). Perception of partly occluded objects in infancy. *Cognitive psychology*, *15*(4), 483–524.

Lyon, D. R., & Waag, W. L. (1995). Time course of visual extrapolation accuracy. *Acta psychologica*, *89*(3), 239–260.

Makin, A. D., & Chauhan, T. (2014). Memory-guided tracking through physical space and feature space. *Journal of Vision*, *14*(13), 10–10.

Makin, A. D., & Poliakoff, E. (2011). Do common systems control eye movements and motion extrapolation? *Quarterly Journal of Experimental Psychology*, *64*(7), 1327–1343.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological review*, *120*(2), 411.

Särkkä, S., & Svensson, L. (2023). *Bayesian filtering and smoothing* (Vol. 17). Cambridge university press.

Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: Clues to visual objecthood. *Cognitive psychology*, *38*(2), 259–290.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420.

Shu, T., Bhandwaldar, A., Gan, C., Smith, K., Liu, S., Gutfreund, D., ... Ullman, T. (2021). Agent: A benchmark for core psychological reasoning. In *International conference on machine learning* (pp. 9614–9625).

Smith, K. A., Dechter, E., Tenenbaum, J. B., & Vul, E. (2013). Physical predictions over time. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).

Smith, K. A., Hamrick, J. B., Sanborn, A. N., Battaglia, P. W., Gerstenberg, T., Ullman, T. D., & Tenenbaum, J. B. (2024). Intuitive physics as probabilistic inference. In T. L. Griffiths, N. Chater, & J. B. Tenenbaum (Eds.), *Bayesian models of cognition : reverse engineering the mind.* Cambridge, MA: MIT Press.

Smith, K. A., Mei, L., Yao, S., Wu, J., Spelke, E., Tenenbaum, J., & Ullman, T. (2019). Modeling expectation violation in intuitive physics with coarse probabilistic object representations. *Advances in neural information processing systems*, *32*.

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in cognitive science*, *5*(1), 185–199.

Tresilian, J. (1995). Perceptual and cognitive processes in time-to-contact estimation: Analysis of prediction-motion and relative judgment tasks. *Perception & Psychophysics*, *57*(2), 231–245.

Vul, E., Alvarez, G., Tenenbaum, J., & Black, M. (2009). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. *Advances in neural information processing systems*, *22*.