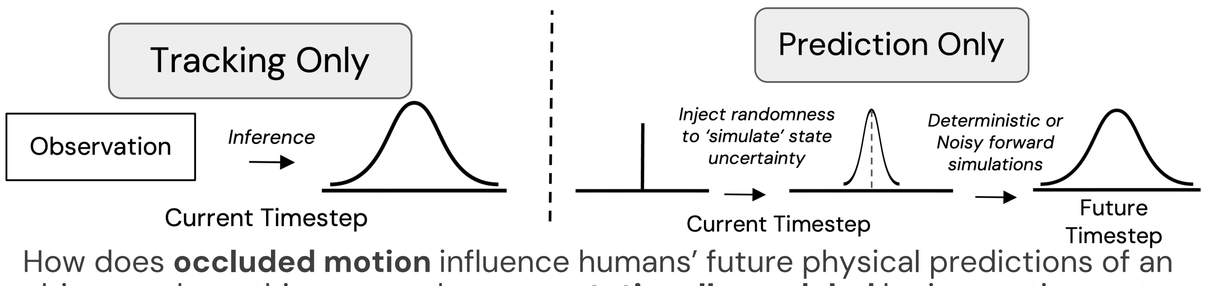
Seeing through Occlusion: Uncertainty-aware Joint Physical Tracking and Prediction

Arijit Dasgupta, Andrew D. Bolton, Vikash K. Mansinghka, Joshua B. Tenenbaum, Kevin A. Smith

Massachusetts Institute of Technology

Background

- Humans can reason about hidden physical processes despite the absence of visually changing evidence.
- Physical tracking and future prediction have traditionally been studied **separately**

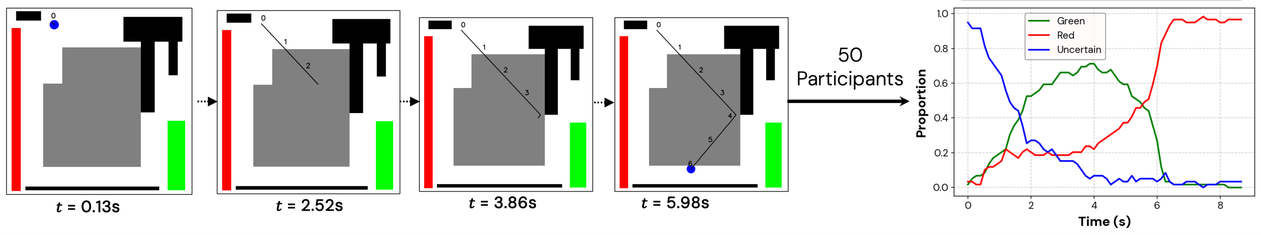


- How does **occluded motion** influence humans' future physical predictions of an object, and can this process be **computationally modeled** by integrating *state estimation* with *prediction*?

Experiment

- **59** participants recruited via Prolific (US$ 15/hr)
- **50** video trials (34 with Occlusion, 16 without) of a ball moving in a 2.5D scene
- **Objective**: continuously predict if the ball will hit red or green first
- Participants are scored between -80 to 120, penalizing mistakes and rewarding correct predictions

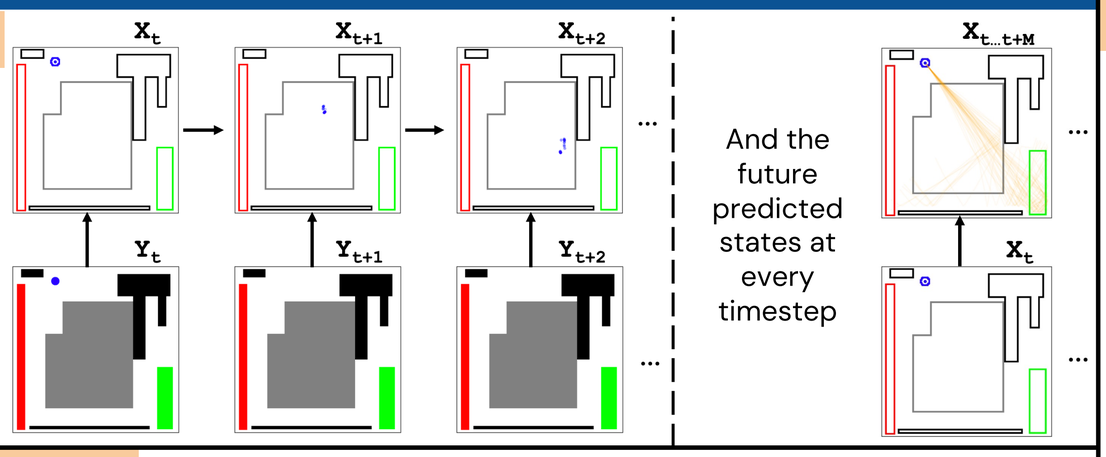
Red-Green Task Press **'F'** for red and '**J**' for green



Joint Tracking and Prediction (JTAP) Model

Objective

Given a stream of images → represent beliefs about the state of the scene



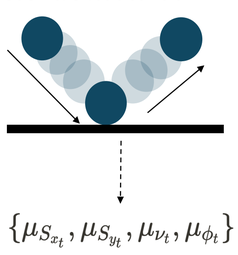
And the future predicted states at every timestep

Probabilistic Model

The state X_t encodes the human observer's mental representation of the object's position, speed and direction

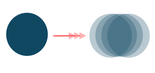
Physical Dynamics Model

Frictionless straight-line motion & elastic collision



$\{\mu_{S_{x_t}}, \mu_{S_{y_t}}, \mu_{\nu_t}, \mu_{\phi_t}\}$

Speed uncertainty

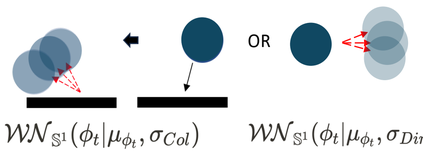


$\mathcal{N}(\nu_t|\mu_{\nu_t}, \sigma_\nu)$

Positional uncertainty

$\mathcal{N}(\mu_{S_{x_t}}, \mu_{S_{y_t}}|\Sigma_S)$

Directional uncertainty



$\mathcal{WN}_{\mathbb{S}^1}(\phi_t|\mu_{\phi_t}, \sigma_{Col})$ $\mathcal{WN}_{\mathbb{S}^1}(\phi_t|\mu_{\phi_t}, \sigma_{Dir})$

Observation Model

$$\hat{Y}_t = \text{Render}(X_t)$$

Static objects → Moving objects →

Occluder → All others →

Red Region → Green Region →

Independent pixel flip noise, used for ***likelihood*** computations

Full Sequential Model:

$$P(X_{0:T}, Y_{0:T}|\rho, \eta) = P(X_0)P(Y_0|X_0, \rho)\prod_{t=1}^{T} P(Y_t|X_t, \rho)P(X_t|X_{t-1}, \eta)$$

Inference Algorithm

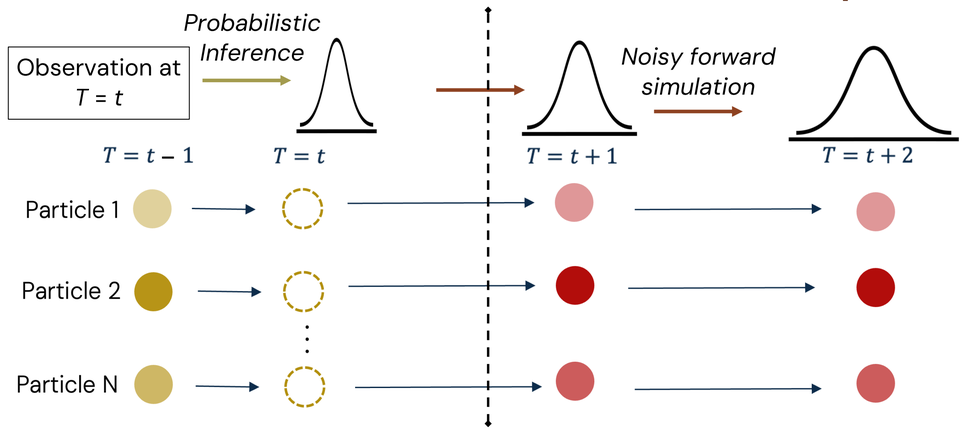
Sequential Monte Carlo algorithm via GPU-accelerated Probabilistic Programming

Implemented in GenJAX

At each timestep

1. Sample Proposal
2. Update weights
3. Resample **all** particles
4. Sample dynamics model ***M*** timesteps

Tracking phase



Future Prediction phase

Proposal

If object is **visible** →use a ***Bottom-up*** proposal to sample position and derived velocity using the observed image

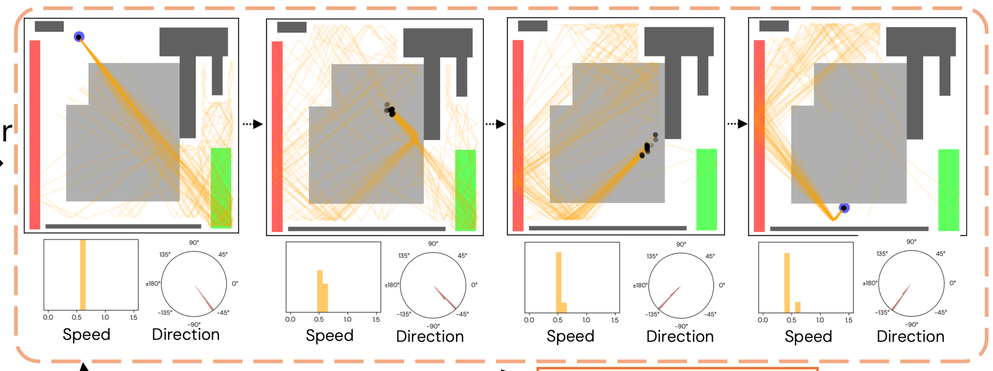
If object is **not visible** → use a ***Top-down*** proposal by sampling from the dynamics model

Results

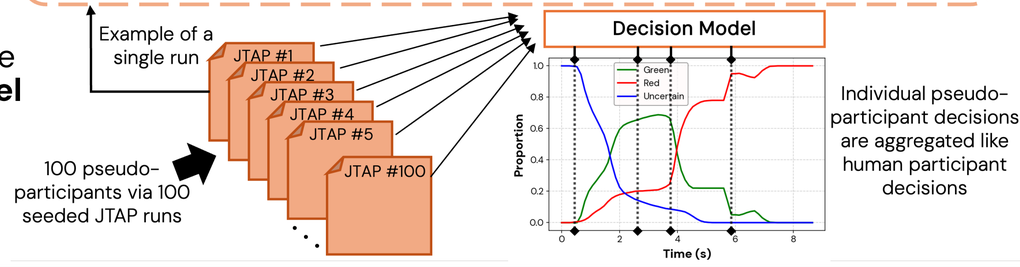
Individual Model Runs

Discrete posterior approximation → View into the "mind" of inference

Single model run proxies the belief states for one pseudo-participant

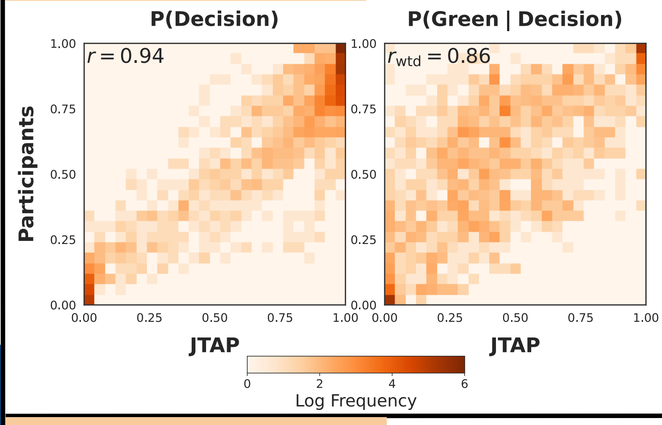


Modeling at the **individual-level** beliefs



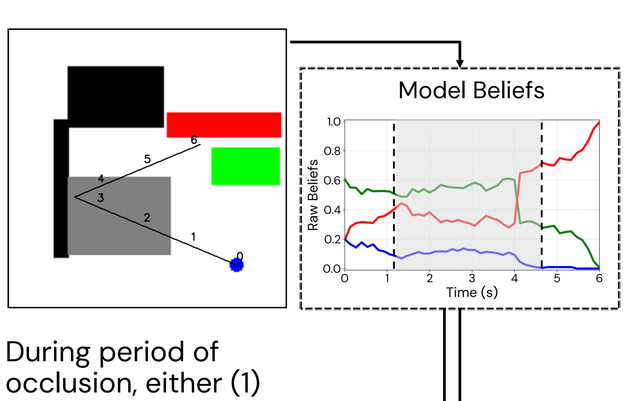
Individual pseudo-participant decisions are aggregated like human participant decisions

Overall Results

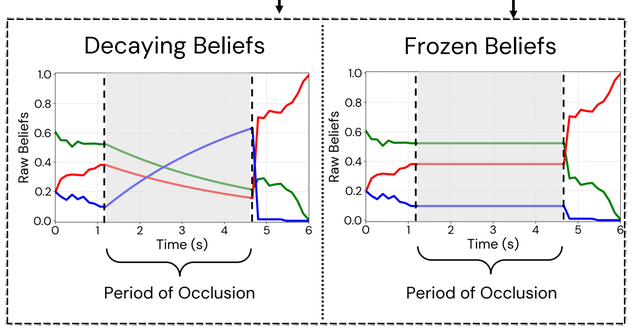


- Participants produces reliable decisions (ICC1k = 0.952)
- JTAP explains well both ***when*** participants decided to push a button (r=0.94) and ***which*** button they pressed (r=0.86)
- Uncertainty gradations between humans are also captured by JTAP

Alternate Models

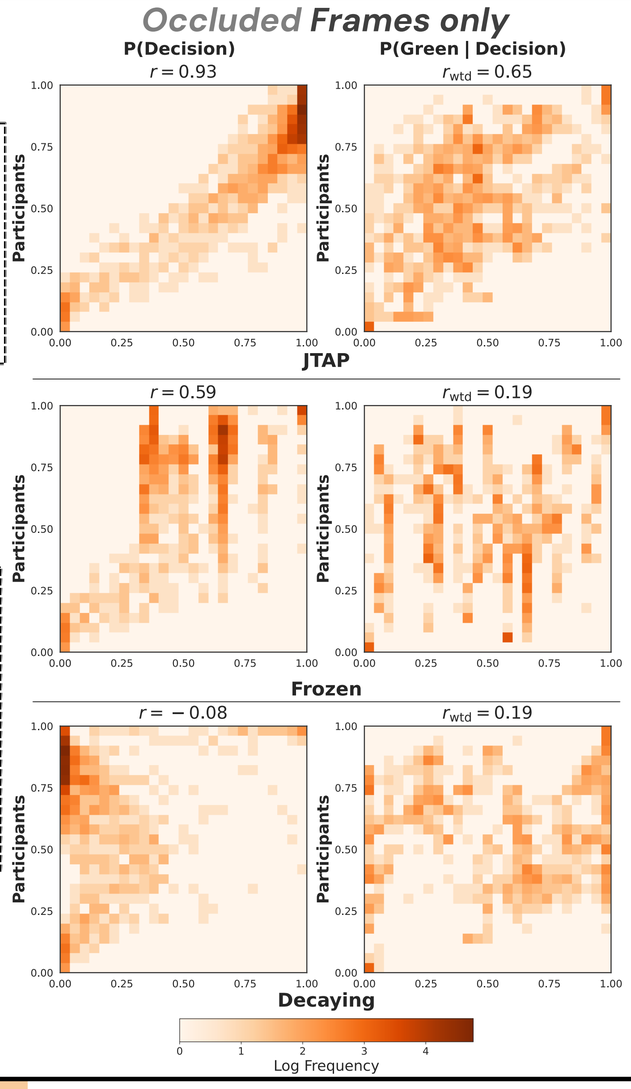


During period of occlusion, either (1) decay the raw beliefs or (2) freeze them

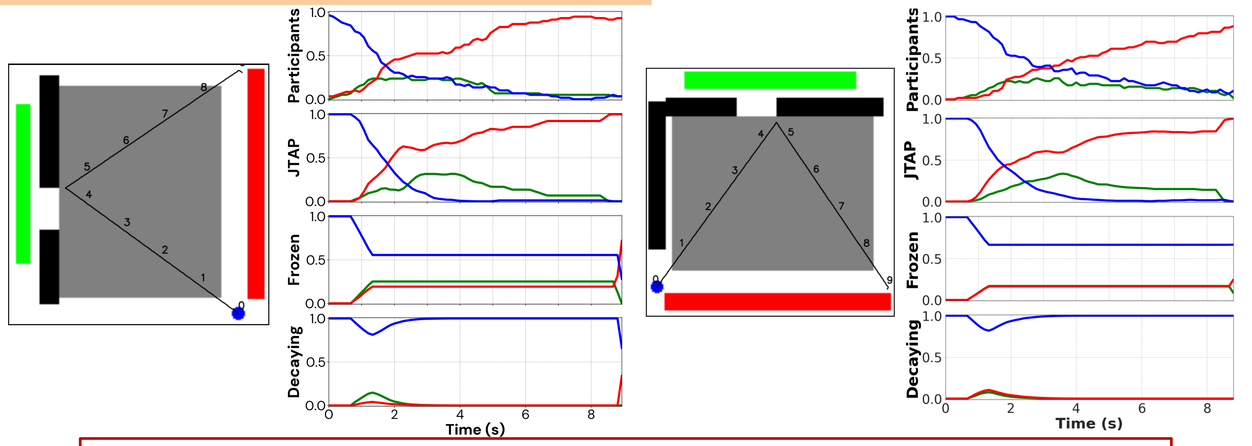


During **occlusion**, JTAP has significantly better correlation with human responses than alternative models ($p < 0.001$)

Occluded Frames only



Benefits of Jointly Tracking and Predicting



The absence of visual evidence over time **is** evidence

Discussion

JTAP models human prediction of object motion during occlusion by integrating **perception**, **probabilistic** reasoning and **physical** knowledge, capturing the rich structure of human physical reasoning.