Decision-Theoretic Learning of Human Perception in Uncertain Environments

Mohammad Alali[1], Tian Lan[2], Seyede Fatemeh Ghoreishi[1], Mahdi Imani[1]
[1]Northeastern University, Boston, MA
[2]George Washington University, Washington, DC
Emails: alali.m@northeastern.edu, tlan@gwu.edu, f.ghoreishi@northeastern.edu, m.imani@northeastern.edu

***Abstract*—Understanding how humans perceive and act in partially known environments is crucial for developing collaborative and assistive systems. In this paper, we introduce a data-driven framework to infer latent perceptual models of humans by representing their movement behavior as a Partially Observable Markov Decision Process (POMDP). We model each possible environment configuration using a discrete set of latent variables and compute the posterior distribution over these configurations based on observed human trajectories. Our approach decomposes the observation likelihood into two components: (I) a prediction model, which is the likelihood of observations given past actions and states, (II) a behavioral model, which is the likelihood of actions given past observations. We model (I) using LSTM networks trained to predict future observations, and we model (II) using a Boltzmann policy based on Q-values learned with Deep Recurrent Q-Networks (DRQN). This allows us to incrementally update our posterior beliefs about the environment in a fully data-driven manner. We validate our method in a simulated rescue maze environment with partially observable and uncertain regions. Extensive experiments show that our framework significantly outperforms the baselines. Our results demonstrate that jointly modeling action and perception yields a powerful tool for understanding human decision-making under uncertainty.**

***Index Terms*—Partially observable Markov decision processes, Deep Reinforcement Learning, Learning from Human Data.**

I. Introduction

Robots and AI assistants must often interpret human behavior in complex, uncertain environments to collaborate effectively. In settings such as assistive robotics, search-and-rescue missions, and autonomous driving, a human's actions are based on *partial and noisy observations* of the world. The ability to infer a person's *latent perceptual model*, i.e., their internal beliefs about the environment, from their behavior is crucial for proactive assistance [1]–[5]. For example, an assistive robot could adjust its strategy if it recognizes that the human is operating under a misconception about the environment. More generally, understanding how humans perceive and respond to partially observable environments enables more robust human–AI coordination. However, inferring latent human knowledge or true hidden states from noisy observation trajectories is challenging: humans exhibit *bounded rationality* and their sensory inputs are noisy, making traditional perception inference methods unreliable in these contexts.

A rich body of research has explored modeling human decision-making and inferring perception/intent from behavior. Inverse reinforcement learning (IRL) methods aim to recover an agent's reward function from demonstrations, but many assume full state observability or perfect environment models [6]–[8]. Some recent efforts have extended IRL to partially observable domains [9]–[11], but these still primarily aim to infer goals or reward structures, not human beliefs. Bayesian inverse planning approaches, such as inverse POMDPs [12], [13] and theory-of-mind models [14], [15], attempt to infer latent mental states by inverting a model of rational behavior. However, these often assume either deterministic observation processes or an equivalence between observations and true states. As a result, they struggle in scenarios where humans operate under perceptual limitations or sensory noise.

Parallel progress in imitation learning, particularly using deep neural networks, has enabled powerful behavioral cloning and adversarial imitation approaches [16]–[20]. These data-driven methods can capture rich human behavior and scale to high-dimensional observations. However, they typically require large training datasets, do not generalize well to unseen environments, and focus on mimicking behavior rather than inferring hidden latent variables. Moreover, standard imitation learning techniques do not explicitly model the agent's observation process, leading to misinterpretation when the demonstrator's behavior results from limited or noisy perception.

In this paper, we propose a novel framework to infer latent human perceptual parameters from behavior in partially observable environments. We model human movement using a POMDP with an unknown latent environment configuration, where each configuration defines a possible observation function and world layout. Our approach jointly models human actions and observations: we use a Deep Recurrent Q-Network (DRQN) [21] to learn a history-dependent Q-function and define a Boltzmann policy to model bounded rationality. Simultaneously, we use LSTMs to learn an observation-prediction model that captures time-series dependencies and sensory uncertainty. This enables a unified, data-driven computation of action and observation likelihoods.

At inference time, we compute the posterior by evaluating the likelihood of the observed trajectory under each environment configuration. Our method updates this belief in an online fashion, allowing it to quickly identify the most probable environment solely based on noisy human observations. Unlike existing approaches, our framework explicitly accounts

for both perceptual uncertainty and decision-making, resulting in more accurate and interpretable model inference.

Our main contributions are:

- We infer latent human perceptual models using only noisy observations via Bayesian inference in a POMDP, combining action and observation modeling.
- We propose a data-driven framework that learns human action policies using DRQN and Boltzmann distributions, and learns observation likelihoods using LSTM-based predictors.
- Our proposed method is implemented as an online inference method to compute posterior distributions over different environment configurations, enabling real-time adaptation and interpretation. We validate our method in a rescue environment, showing improved inference accuracy over two baselines that ignore observation noise or perceptual modeling.

II. Background

Human movement in physical environments can be rigorously modeled using the framework of *Partially Observable Markov Decision Processes* (POMDPs), which formalize sequential decision-making under uncertainty [22]. A POMDP is defined by the tuple $(S, A, Z, T, O, R, \gamma)$, where S denotes the set of latent environment states, A the set of actions available to the human, and Z the set of possible observations. The state transition function $T(s' \mid s, a, \theta)$ gives the probability of moving to state $s' \in S$ after taking action $a \in A$ in state s, under a latent perceptual parameter $\theta \in \Theta$. The observation function $O(z \mid s', a, \theta)$ specifies the probability of receiving observation $z \in Z$ after executing action a and transitioning to state s'. The reward function $R^\theta(s, a)$ assigns a scalar value to taking action a in state s, and the discount factor $\gamma \in [0, 1)$ encodes the human's preference for immediate versus future rewards. This formulation provides a principled structure for capturing how humans act, perceive, and make decisions in uncertain environments.

A key aspect of our formulation is the introduction of a latent perceptual parameter $\theta \in \Theta$, representing the human's internal interpretation of the environment. We assume a finite set of candidate perceptual models,

$$\Theta = \{\theta^1, \theta^2, \ldots, \theta^N\}, \tag{1}$$

where each θ parametrizes a distinct observation model and captures differences in sensory processing, attention, or prior knowledge. Under this formulation, the observation function becomes

$$O(z \mid s', a, \theta) : Z \times S \times \mathcal{A} \times \Theta \to [0, 1], \tag{2}$$

allowing each environment—or, more precisely, each human's perception of it—to be governed by a specific θ. Treating perception as a latent variable enables personalized inference and supports reasoning over heterogeneous behavioral patterns.

III. Problem Formulation

Our primary objective is to infer the latent perceptual parameter $\theta \in \Theta$ solely from a sequence of observed data $Z_{1:k} = \{Z_1, Z_2, \ldots, Z_k\}$. By learning the posterior distribution over θ, we gain insight into human perception and, by extension, their possible future actions.

To represent the full posterior over all candidate perceptual parameters in $\Theta = \{\theta^1, \theta^2, \ldots, \theta^N\}$, we define a probability vector $P_k \in \mathbb{R}^N$ as:

$$P_k := \Big[P(\theta^1 \mid Z_{1:k}), P(\theta^2 \mid Z_{1:k}), \ldots, P(\theta^N \mid Z_{1:k})\Big]. \tag{3}$$

where $\sum_{i=1}^{N} P(\theta^i \mid Z_{1:k}) = 1$. Each element of this vector quantifies the likelihood that a specific perception model θ^i best explains the sequence of observed data.

We aim to compute the posterior:

$$P(\theta \mid Z_{1:k}) \propto P(Z_{1:k} \mid \theta) P(\theta). \tag{4}$$

To compute $P(\theta \mid Z_{1:k})$, we expand the likelihood term using the chain rule:

$$\begin{aligned} P(Z_{1:k} \mid \theta) &= \prod_{r=2}^{k} P(Z_r \mid Z_{1:r-1}, \theta) \\ &= \prod_{r=2}^{k} \sum_{a \in \mathcal{A}} P(Z_r, a_{r-1} = a \mid Z_{1:r-1}, \theta) \\ &= \prod_{r=2}^{k} \sum_{a \in \mathcal{A}} \underbrace{P(Z_r \mid a_{r-1} = a, Z_{1:r-1}, \theta)}_{\text{(I) Prediction Model}} \\ &\qquad \underbrace{P(a_{r-1} = a \mid Z_{1:r-1}, \theta)}_{\text{(II) Behavioral Model}}, \end{aligned} \tag{5}$$

This decomposition expresses the joint likelihood in terms of:

- **(I) Prediction Model**: the probability of receiving observation Z_r conditioned on the previous action $a_{r-1} = a$, prior observations $Z_{1:r-1}$, and perceptual parameter θ,
- **(II) Behavioral Model**: the probability of the human taking action a_{r-1} given prior observations and θ.

Note that both (I) and (II) implicitly incorporate all sources of observation noise and behavioral stochasticity. These components model the inherent uncertainty in sensing and movement during human navigation. There is no closed-form or conventional analytical solution to estimate (I) and (II), nor to update the posterior $P(\theta \mid Z_{1:k})$ incrementally as new observations arrive. In this work, we tackle this challenging problem directly. In the following sections, we present our proposed data-driven approach to learn the necessary probabilistic models for (I) and (II), enabling scalable and accurate inference of $P(\theta \mid Z_{1:k})$ from raw observational sequences.

IV. Learning Human Perception and Decision Processes

A. Behavioral Model

To formalize human decision-making, the history-based human policy is defined, accounting for all past observations that shape decision-making. The history space at time step k is defined as $h_k = \{Z_{1:k}\} \in \mathbb{R}^{d\times k}$, where $Z_i \in \mathbb{R}^d$ and each history $h_k \in \mathcal{H}_k$ consists of all past observations up to time k, forming a growing sequence that encapsulates the evolving human perception of the environment. Given this structure, a stationary human policy is formulated as a mapping $\pi_\theta : \mathcal{H} \to \mathcal{A}$, where $\pi_\theta(h)$ determines the probability of selecting an action based on the full history of observations. The optimal history-based policy for the human is then expressed as

$$\pi^*_\theta(h) := \arg\max_{\pi_\theta} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t R^\theta(s_t, a_t) \,\middle|\, h_0 = h, a_t \sim \pi_\theta(\cdot \mid h_t)\right], \tag{6}$$

where $\gamma \in (0, 1]$ is the discount factor, T is the horizon, and $\mathbb{E}[\cdot]$ is taken over the trajectory of future states and actions under policy π_θ. The state s_t is only partially observable; hence, the history h_t serves as a sufficient statistic representing the human's internal belief about the state. However, the computation of the optimal policy π^*_θ is generally intractable due to the continuity and high dimensionality of the history space.

Correspondingly, we define the optimal action-value function $Q^*_\theta(h, a)$ as the expected discounted return starting from history h, taking action a, and thereafter following the optimal policy π^*:

$$Q^*_\theta(h,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R^\theta(s_t, a_t) \,\middle|\, h_0 = h, a_0 = a, a_{t>0} \sim \pi^*_\theta(\cdot \mid h_t)\right]. \tag{7}$$

In our framework, we model the human decision-making policy $\pi^H_\theta(a \mid h)$ under perceptual parameter θ as a *Boltzmann (softmax) policy* based on the optimal Q-function $Q^*_\theta(h, a)$. Specifically:

$$\pi^H_\theta(a \mid h) = \frac{\exp\left(\eta Q^*_\theta(h, a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(\eta Q^*_\theta(h, a')\right)}. \tag{8}$$

Here, $\eta \in \mathbb{R}^+$ is the temperature parameter controlling the stochasticity of the policy: higher values of η result in more deterministic (greedy) policies, while lower values induce more exploratory behavior.

Given the sequential nature of observations h_k and the discrete action space $\mathcal{A}$, we use a deep reinforcement learning technique called *Deep Recurrent Q-Network (DRQN)* [21] to approximate the optimal action-value function $Q^*_\theta(h, a)$ for each environment θ. DRQN is well-suited for partially observable domains as it maintains a recurrent hidden state that captures temporal dependencies in observation histories. Using the optimal Q-function $Q^*_\theta(h, a)$ learned via DRQN and the human policy in (8), we can directly compute the likelihood of observing a human action given a history of observations. Thus, term (II) in (5) is computed as:

$$P(a_{r-1} = a \mid Z_{1:r-1}, \theta) = \pi^H_\theta(a \mid h_{r-1}), \tag{9}$$

where the Q-function $Q^*_\theta(h, a)$ is trained using DRQN for each environment θ.

B. Likelihood Model and Neural Network Approximation

We now turn to estimating term (I) in Equation (5), namely the likelihood of observing Z_r given the past observations $Z_{1:r-1}$, the previous action a_{r-1}, and θ. To capture the uncertainty in observations, the likelihood function is commonly modeled using a parametric distribution, such as a Gaussian, where the mean represents the expected observation and the variance encodes uncertainty. Formally, the likelihood function can be defined as

$$P(Z_r | a_{r-1} = a, Z_{1:r-1}, \theta) \propto \exp\left(-\frac{\|\mathbb{E}[Z_r | a_{r-1}, Z_{1:r-1}, \theta] - Z_r\|_2^2}{\tau}\right), \tag{10}$$

where $E[Z_r | a_{r-1}, Z_{1:r-1}, \theta]$ represents the expected observation under model θ, and τ is a temperature parameter controlling the sharpness of the distribution. This Gaussian-based formulation assumes that deviations from the expected observation follow an isotropic noise model, making it particularly effective when observation noise is roughly homogeneous and independent.

We employ a neural network to approximate the prediction model $P(Z_r | a_{r-1}, Z_{1:r-1}, \theta)$. Let $F^\theta_W(Z_{1:r-1}, a_{r-1})$ represent a recurrent neural network (e.g., LSTM) that takes as input a sequence of past observations and the most recent action and outputs an estimate of the expected next observation $E[Z_r | Z_{1:r}, a_{r-1}, \theta]$ under model θ. Given that the trajectories are generated from the model θ under the human policy π^H_θ, the network is trained to minimize the following loss function:

$$\mathcal{L}_{\text{MSE}} = \left\| \bar{Z}_r - F^\theta_W(Z_{1:r-1}, a_{r-1}) \right\|_2^2, \tag{11}$$

where $\bar{Z}_r$ is the true observation at time step r. While we use LSTM in this work due to its established success in time-series modeling, other architectures such as Transformer-based models could also be used to model observation dynamics.

Once trained, the neural network is used to define the approximate likelihood function for each model θ. Instead of computing $P(Z_r | a_{r-1}, Z_{1:r-1}, \theta)$ analytically, the likelihood is modeled using an exponential weighting function:

$$P(Z_r \mid a_{r-1}, Z_{1:r-1}, \theta) = \frac{\exp\left(-\xi \left\| \bar{Z}_r - F^\theta_W(Z_{1:r-1}, a_{r-1}) \right\|_2\right)}{\sum_{a' \in \mathcal{A}} \exp\left(-\xi \left\| \bar{Z}_r - F^\theta_W(Z_{1:r-1}, a') \right\|_2\right)}. \tag{12}$$

A larger ξ leads to a more deterministic likelihood distribution, assigning higher probability to models whose predicted observation is closer to the actual observation. A smaller ξ introduces more stochasticity, broadening the likelihood distribution over possible models.

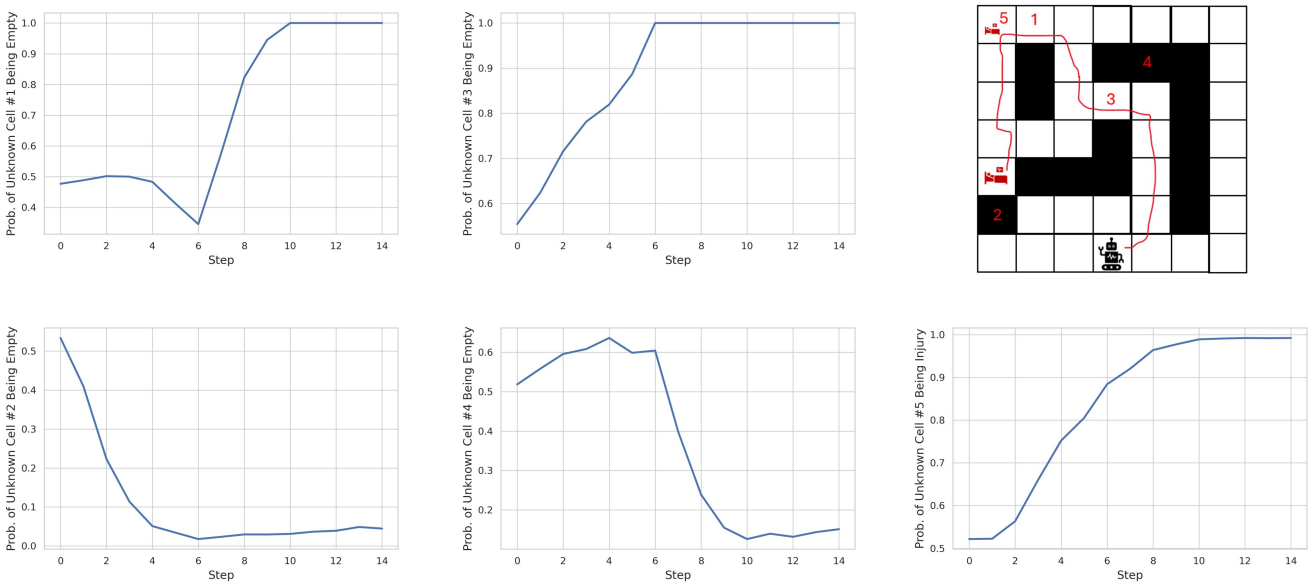

Fig. 1: Evolution of posterior probabilities for uncertain cells over time based on a sample observation trajectory.

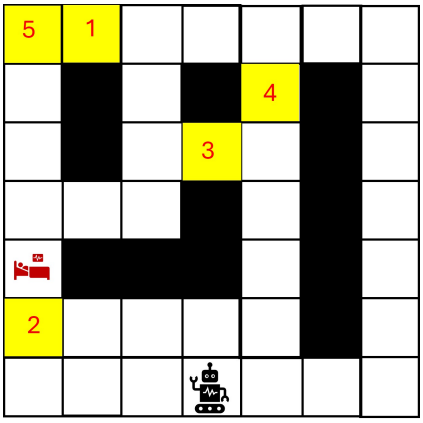

Fig. 2: Rescue maze environment used in experiments. Cells can be *empty* (white), *walls* (black), or *injuries* (red hospital icon). Yellow-labeled cells represent uncertain elements of the environment.

C. Human Decision and Perception Posterior

Having defined the models and training schemes for both components, we are now able to compute terms (I) and (II) in (5) using (12) and (9), respectively. This enables us to estimate the posterior distribution $P(\theta \mid Z_{1:k})$ in a fully data-driven manner at each time step, incrementally updating our belief about the underlying human perception as new observations become available. The posterior distribution over human perception is given by

$$\begin{aligned} P_k(\theta) &\propto \\ &\left[\prod_{r=2}^{k} \sum_{a\in\mathcal{A}} P(Z_r|a_{r-1}=a, Z_{1:r-1}, \theta) P(a_{r-1}=a|Z_{1:r-1}, \theta)\right] P(\theta) \\ &= \left[\prod_{r=2}^{k} \sum_{a\in\mathcal{A}} \vartheta^{\theta}(Z_r|Z_{1:r-1}, a)\pi_{\theta}^{H}(a \mid h)\right] P(\theta), \end{aligned} \tag{13}$$

where $\vartheta^{\theta}(Z_r|Z_{1:r-1}, a)$ represents the likelihood model approximated using a trained LSTM-based neural network, and $\pi_{\theta}^{H}(a \mid h)$ characterizes the human behavioral model, inferred through reinforcement learning techniques.

Given an observed sequence of measurements $Z_{1:k}$, the probability of a human selecting action a_k at time step k follows

$$P(a_k = a \mid Z_{1:k}) = \sum_{\theta\in\Theta} P(a_k, \theta \mid Z_{1:k}) = \sum_{\theta\in\Theta} \pi_{\theta}^{H}(a \mid h) P_k(\theta).$$

Here, $P_k(\theta)$ represents the posterior belief over the human's internal model parameters, reflecting the uncertainty in human perception based on the observed trajectory. Unlike traditional reinforcement learning frameworks that assume a fully known reward function and transition model, human decision-making under uncertainty is influenced by the inferred action-value function $Q_{\theta}^{*}(Z_{1:k}, a)$, which quantifies the expected long-term utility of selecting action a_k given the observation history under model θ.

This formulation provides a robust framework for predicting human decisions while simultaneously refining the underlying belief distribution over their perceived environment. By leveraging data-driven approximations through DRQN and LSTM-based networks, this approach enables scalable inference, capturing complex decision-making dynamics that would otherwise be infeasible to model explicitly.

V. Numerical Experiments

We evaluate our approach in a simulated rescue mission within a grid-based maze (Fig. 2), where a human-controlled robot navigates the environment to locate injured individuals. Each cell is either *empty*, a *wall* (impassable), or an *injury* location. A subset of cells is intentionally uncertain: cells 1–4 (highlighted in yellow) may be *empty* or *wall*, and cell 5 may be *empty* or an *injury*. These ambiguous cell states define the

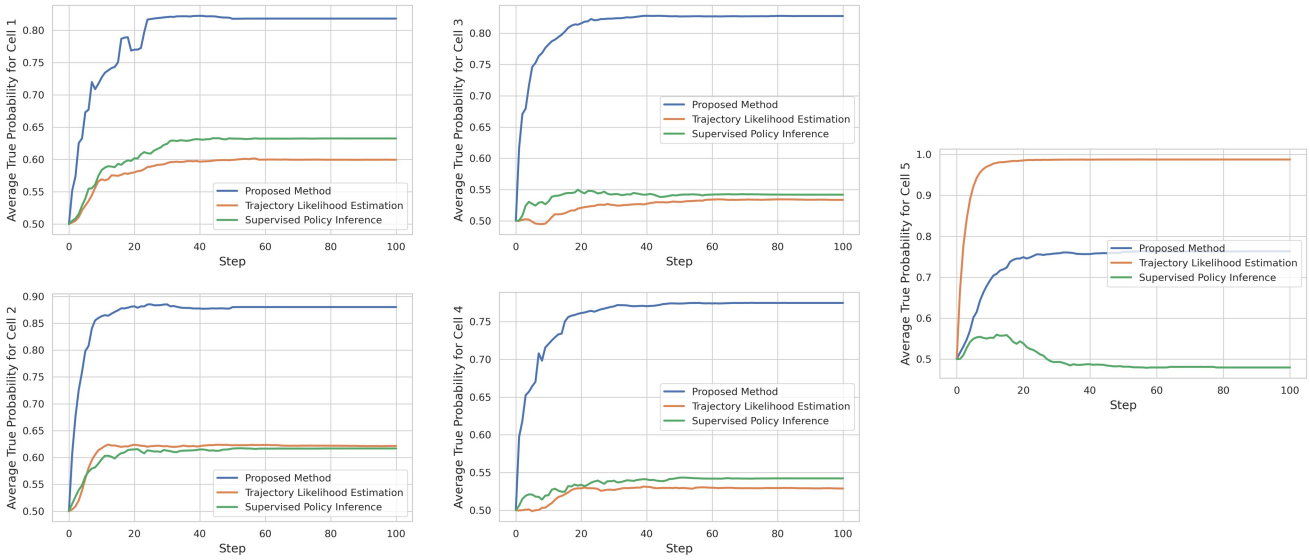

Fig. 3: Average true probability for each uncertain cell across 100 steps, comparing the proposed method with two baseline approaches.

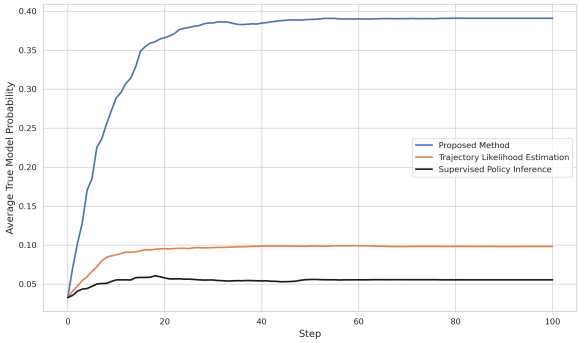

Fig. 4: Average true model probability over 100 steps, comparing the proposed method against baseline approaches across all environments.

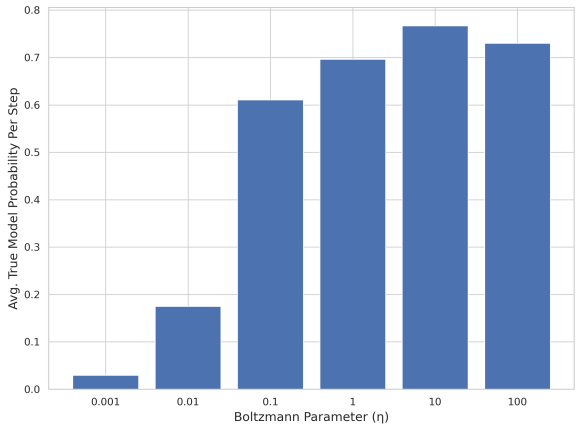

Fig. 5: Ablation study on Boltzmann parameter η: average true model probability per step.

latent environment configuration, yielding $2^5 = 32$ possible instances.

The human can move up, down, left, or right, with 90% probability of executing the intended direction and 5% probability of drifting to either perpendicular direction. Observations are also noisy: the agent's true position is reported with 90% probability and otherwise replaced uniformly by one of the adjacent cells. A positive reward is given for visiting an injury cell, and the goal is to infer the latent environment configuration θ from human trajectories so that future agents can rapidly identify injuries. To apply our framework, we enumerate all 32 possible configurations and, for each θ, independently train a DRQN-based human policy (using a Boltzmann action model) and an LSTM predictor for observation dynamics. These models enable data-driven computation of $P(\theta \mid Z_{1:k})$ throughout the trajectory via (5). In our first experiment, the ground-truth environment contains empty cells at positions 1 and 3, walls at positions 2 and 4, and an injury at position 5. Using this configuration, we generate a 14-step human observation trajectory (shown in the top-right of Fig. 1) and compute the posterior $P(\theta \mid Z_{1:k})$ at each time step.

To visualize how the agent's beliefs evolve, we marginalize the posterior over environments to obtain per-cell probabilities by summing the posterior mass of all configurations consistent with each cell state (e.g., empty or injury). Figure 1 shows the resulting per-cell posteriors over time. As the agent approaches cell 3, the probability of it being empty rises sharply to 1, while the probability of cell 2 being empty decreases as the agent moves away from the known wall. Similar updates occur near cells 1 and 5, where the model increasingly favors cell 1 being empty and cell 5 being an injury. These spatial–temporal trends demonstrate that the framework effectively integrates noisy observations to refine beliefs about individual cells and, more broadly, about the underlying environment hypothesis.

In the next experiment, we assess the robustness and generalization of our approach across all 32 environment configurations. For each configuration, we generate 10 random observation trajectories and compute the posterior $P(\theta \mid Z_{1:k})$ over time, averaging the results across all runs. We com-

pare our method to two baselines: (i) Trajectory Likelihood Estimation, which relies solely on observation sequences and ignores human perception, and (ii) Supervised Policy Inference, which trains a classifier to map full trajectories to discrete environment identities. As shown in Fig. 4, our method achieves substantially higher true-model probabilities over time than both baselines. Figure 3 further shows the average per-cell accuracy for the five uncertain cells, where our method outperforms the alternatives on nearly all cells, with a slight exception for cell 5. Overall, the results demonstrate that explicitly modeling human perception yields markedly superior inference performance at both the model and cell levels.

In our final experiment, we conduct an ablation study on the Boltzmann parameter η, which controls the sharpness of the human policy $\pi_\theta^H(a \mid h)$. We evaluate six values $\eta \in \{0.001, 0.01, 0.1, 1, 10, 100\}$ by generating 10 trajectories from each of the 32 environments and computing the average true-model probability per step. As shown in Fig. 5, very small η values yield poor performance due to near-uniform action selection, while larger values sharpen the policy and improve discrimination between environments. Performance peaks at $\eta = 10$, with a slight drop at $\eta = 100$, suggesting overconfident determinism that reduces robustness to noise. These results underscore the importance of properly calibrating η to balance expressiveness and stability in behavior-based inference.

VI. Conclusion

We introduced a data-driven framework for inferring latent human perceptual parameters by modeling human movement as a Partially Observable Markov Decision Process (POMDP). Our approach combines action and observation likelihoods in a unified Bayesian inference procedure over a discrete set of candidate environments, using a Deep Recurrent Q-Network with a Boltzmann policy to model action selection and an LSTM-based predictor to capture observation dynamics under noise and partial observability. In grid-based rescue experiments, the method accurately recovered underlying environment configurations and outperformed baseline approaches in both model identification and per-cell inference. An ablation on the Boltzmann temperature further highlights the role of policy sharpness in successful perception inference. Overall, the framework provides a principled and generalizable way to interpret human behavior in uncertain environments, with applications to assistive robotics, collaborative planning, and human-in-the-loop systems.

Acknowledgment

The authors acknowledge the support of the National Science Foundation award IIS-2311969, ARMY Research Laboratory awards W911NF-23-2-0207 and W911NF-24-2-0166, and Office of Naval Research award N00014-23-1-2850.

References

[1] D. S. Brown and S. Niekum, "Better imitation learning through principled regularization," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 3715–3722, 2020.

[2] S. Arora and P. Doshi, "A survey of inverse reinforcement learning: Challenges, methods and progress," *Artificial Intelligence*, vol. 297, p. 103500, 2021.

[3] N. Asadi, S. H. Hosseini, M. Imani, D. P. Aldrich, and S. F. Ghoreishi, "Privacy-preserved federated reinforcement learning for autonomy in signalized intersections," in *International Conference on Transportation and Development 2024*, pp. 390–403, 2024.

[4] N. Asadi and S. F. Ghoreishi, "Efficient learning of uncertainty distributions in coupled multidisciplinary systems through sensory data," *IET Cyber-Physical Systems: Theory & Applications*, vol. 10, no. 1, p. e70000, 2025.

[5] Y. Lin, S. F. Ghoreishi, T. Lan, and M. Imani, "Reinforcement Learning for Human-AI Collaboration via Probabilistic Intent Inference," in *Proceedings of the Reinforcement Learning Conference (RLC)*, 2025.

[6] Y. Wang, P. Wu, and M. Imani, "Federated Posterior Sharing for Multi-Agent Systems in Uncertain Environments," in *7th Annual Learning for Dynamics & Control Conference*, PMLR, 2025.

[7] A. KazemiNajafabadi, M. Everett, T. Lan, N. Bastian, and M. Imani, "Adversarial decoy placement for strategic state perturbations in artificial intelligence driven defense," in *Proceedings of the 64th IEEE Conference on Decision and Control (CDC)*, 2025.

[8] A. Ravari, S. F. Ghoreishi, T. Lan, N. Bastian, and M. Imani, "Hybrid Modeling of Heterogeneous Human Teams for Collaborative Decision Processes," in *7th Annual Learning for Dynamics & Control Conference*, PMLR, 2025.

[9] A. X. Djeumou, T. Hoang, C. Reardon, and P. Doshi, "Inverse reinforcement learning with partially observable environment dynamics," in *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pp. 1292–1300, 2021.

[10] A. Ravari, S. F. Ghoreishi, and M. Imani, "Optimal inference of hidden Markov models through expert-acquired data," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 8, pp. 3985–4000, 2024.

[11] D. Wang and S. F. Ghoreishi, "Robust reinforcement learning for autonomous driving in uncertain environments," in *2025 IEEE 21st International Conference on Automation Science and Engineering (CASE)*, pp. 1436–1443, IEEE, 2025.

[12] Y. Lin, S. F. Ghoreishi, T. Lan, and M. Imani, "High-level human intention learning for cooperative decision-making," in *2024 IEEE Conference on Control Technology and Applications (CCTA)*, pp. 209–216, IEEE, 2024.

[13] Y. Wu, Y. Tian, Y. Zhu, and S.-C. Zhu, "Inverse pomdp: Inferring what you think about what i think," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 823–833, 2018.

[14] C. Baker, J. Jara-Ettinger, R. Saxe, and J. B. Tenenbaum, "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing," *Nature Human Behaviour*, vol. 1, no. 4, pp. 1–10, 2017.

[15] M. Alali and M. Imani, "Deep Reinforcement Learning Data Collection for Bayesian Inference of Hidden Markov Models," *IEEE Transactions on Artificial Intelligence*, 2024.

[16] J. Ho and S. Ermon, "Generative adversarial imitation learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4565–4573, 2016.

[17] R. Rafailov, A. Rajeswaran, C. Finn, and S. Levine, "Visual imitation with latent environment inference," in *Conference on Robot Learning (CoRL)*, pp. 1331–1340, 2021.

[18] D. Wang and S. F. Ghoreishi, "Rgdr: Reward-guided domain randomization for autonomous driving," in *2025 IEEE 28th International Conference on Intelligent Transportation Systems (ITSC 2025)*, IEEE, 2025.

[19] A. Kazeminajafabadi and M. Imani, "Optimal monitoring and attack detection of networks modeled by Bayesian attack graphs," *Cybersecurity*, vol. 6, no. 1, p. 22, 2023.

[20] A. Kazeminajafabadi, T. Lan, and M. Imani, "Game-theoretic defense policy for network security against intelligent adversary," in *21st IEEE International Conference on Automation Science and Engineering (CASE)*, IEEE, 2025.

[21] M. Hausknecht and P. Stone, "Deep recurrent q-learning for partially observable mdps," in *2015 AAAI Fall Symposium Series*, 2015.

[22] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial intelligence*, vol. 101, no. 1-2, pp. 99–134, 1998.