ENHANCING HUMAN-AI COLLABORATION THROUGH LOGIC-GUIDED REASONING

Chengzhi Cao[1,2]*, **Yinghao Fu**[1]*, **Sheng Xu**[1], **Ruimao Zhang**[1], **Shuang Li**[1]†
[1]The Chinese University of Hong Kong (Shenzhen)
[2]University of Science and Technology of China
chengzhicao@mail.ustc.edu.cn, yinghaofu@link.cuhk.edu.cn,
shengxu1@link.cuhk.edu.cn, zhangruimao@cuhk.edu.cn
lishuang@cuhk.edu.cn

ABSTRACT

We present a systematic framework designed to enhance human-robot perception and collaboration through the integration of logical rules and Theory of Mind (ToM). Logical rules provide interpretable predictions and generalize well across diverse tasks, making them valuable for learning and decision-making. Leveraging the ToM for understanding others' mental states, our approach facilitates effective collaboration. In this paper, we employ logic rules derived from observational data to infer human goals and guide human-like agents. These rules are treated as latent variables, and a rule encoder is trained alongside a multi-agent system in the robot's mind. We assess the posterior distribution of latent rules using learned embeddings, representing entities and relations. Confidence scores for each rule indicate their consistency with observed data. Then, we employ a hierarchical reinforcement learning model with ToM to plan robot actions for assisting humans. Extensive experiments validate each component of our framework, and results on multiple benchmarks demonstrate that our model outperforms the majority of existing approaches.

1 INTRODUCTION

Humans are natural mind-readers. The capacity to deduce what others are thinking or wanting from brief nonverbal interactions, called Theory of Mind (ToM), is crucial to our social lives (Mao et al., 2023; Fuchs et al., 2023).We know not only how to take actions to interact with the world but, moreover, how to *purposefully* influence the thoughts of other humans (collaborators or opponents) to achieve our goal (Chandra et al., 2020). This intellectual ability enables humans to work smoothly and strategically with other humans, which distinguishes humans from other lower-level animals. An intriguing question arises:

Could we endow AI agents with this ability so that the AI agents can be more intellectually human-like?

Particularly in a human-AI collaborative task setting, if AI agents can adeptly reason about the unobservable mental states of humans, they could engage more effectively in cooperative tasks, facilitating seamless human-AI interactions (Rabinowitz et al., 2018; Grassiotto & Costa, 2021). For example, a self-driving car equipped with the ToM can better communicate with other human drivers nonverbally. Faced with sophisticated or ambiguous traffic scenarios, ToM enables self-driving cars to reason about the mental states, specifically the beliefs, of human drivers sharing the road (Patel & Chernova, 2022). For another instance, suppose one day we aim to develop an AI agent to aid and take care of elder people, it requires understanding human nonverbal communication, such as hand gestures, or eye contact, to understand the intentions or desires of humans. An agent equipped with ToM is endowed with the ability to interpret these cues more effectively.

*Equal Contribution
†Corresponding author

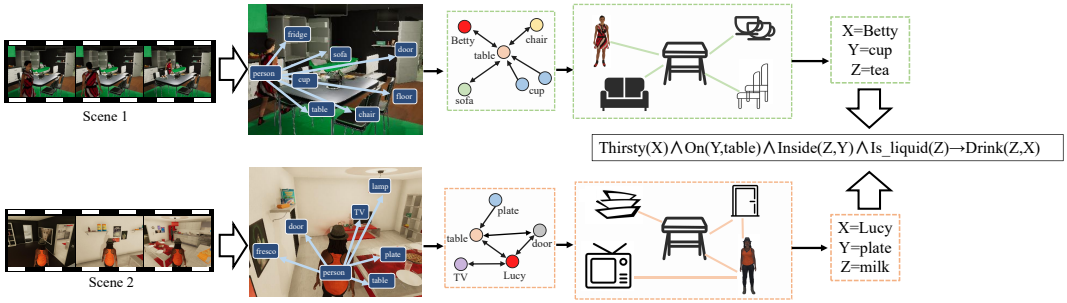


Figure 1: Using logical rules (shown in the black box) for human action perception.

To achieve the level of social intelligence, the robots must develop two key skills in order to help humans effectively: 1) comprehend human behavior (called social perception); 2) reason critically about the environment and organize its actions to work with humans (Puig et al., 2020). Researchers typically refer to the two agents engaged in behavior modeling as "actor" and "observer". The actor behaves in some way based on his own perception of the world. The observer watches the actor and forms an understanding of the mental state held by the actor and/or the resulting actions that the actor intends to take. In the simplest case, the actor behaves deterministically, and the observer has full knowledge of the world external to the actor (Wei et al., 2018). In more complex cases, observers can have partial knowledge, and agents can be both actors and observers at the same time, simultaneously modeling each other.

Complex social movements often have underlying logic that is clear and generalizable. Logic rules serve as high-level representations for reasoning, and guiding actions in specific scenarios (Duan et al., 2022; Yang et al., 2017). These interpretable rules not only aid in understanding inference outcomes but also enhance robustness in diverse situations. The burgeoning field of neurosymbolic AI exemplifies this trend, merging neural networks with Symbolic AI to foster robust and adaptable models. It maintains the validity of logic rules across various contexts without retraining, contrasting with methods reliant on specific entity embeddings (Kautz, 2022; d'Avila Garcez & Lamb, 2020). Figure 1 illustrates an example of rules. If new facts about more scenes or locations are added to the knowledge base, the rule about "Drink(person,water)" will still be usefully accurate without retraining. The same might not be true for methods that learn embeddings for specific knowledge base entities.

In this paper, we leverage probabilistic logical reasoning to generate generalized knowledge about an agent's goal in new environments, enhancing social perception and human-AI collaboration. This approach, grounded in probabilistic logic, is adept at handling uncertainties and modeling complex data. Our methodology intertwines perception and collaboration learning, where perception informs logic rule learning, and logical reasoning guides collaborative efforts. This reciprocal enhancement offers two key advantages: (*i*) **Rule-based Probabilistic Social Perception**: We employ a rule-based probabilistic model, treating each hidden rule as a flexible constraint. This approach accommodates uncertainties, with rules executed based on probabilistic assessments. The rules capture both object relations and event sequence constraints. (*ii*) **Dynamic Collaboration**: Our human-like agent adapts its subgoals and beliefs in real-time, guided by the latest observations and logical rules. It formulates and executes concurrent plans alongside other agents, ensuring robust task performance in partially observable environments and generating human-like behaviors.

2 Related Work

Behavior understanding. The task of recognizing and interpreting human activities faces significant challenges due to the variability in execution styles and the complexity of visual stimuli. Previous research primarily centered on visual behavior modeling (Chen et al., 2021) and trajectory prediction (Cao et al., 2023; Gupta et al., 2018; Alahi et al., 2016; Chandra et al., 2020), with some studies focusing on group activities and plans (Ibrahim et al., 2016; Shu et al., 2015; Choi & Savarese, 2013). Our approach diverges from these works by focusing on activity understanding that infers mental states (e.g., intentions, desires) from behaviors, aligning closely with computational

Theory of Mind (Jara-Ettinger, 2019; Rabinowitz et al., 2018; Nguyen et al., 2022). We aim to synergize behavior understanding with logical reasoning, enhancing the machine's comprehension of human actions through a unified, interactive framework.

Logical reasoning. The integration of logical reasoning with deep learning architectures has garnered considerable interest (Kautz, 2022; d'Avila Garcez & Lamb, 2020). Most existing systems develop differentiable modules based on known relationships for deep networks to learn (Zhou, 2019; Mao et al., 2023; Evans & Grefenstette, 2018). Innovations in this field include learning numerical rules from knowledge graphs (Wang et al., 2019b), differentiable maximum satisfiability solvers (Wang et al., 2019a), and neural-symbolic reasoning models (Tang et al., 2022). However, these models often simplify relations and rely on basic triplet-based embeddings. Our approach advances this field by embedding more intricate first-order logical rules and jointly modeling entities, relations, and logic rules.

Human-robot interaction. Historically, human-robot interaction studies have been confined to controlled lab settings (Goodrich & Schultz, 2007; Dautenhahn, 2007). Notable developments include frameworks for learning from joint-action demonstrations (Nikolaidis et al., 2015) and modulating robot behavior based on user actions (Rozo et al., 2016). Other approaches, like natural language interaction (Gervasi et al., 2020; Ogenyi et al., 2021), focus on creating user models within a Partially Observable Markov Decision Process (POMDP) framework. Our research departs from these methods by utilizing logic rules to infer dominant human goals and efficiently associate new users with types in diverse environments, streamlining the learning process.

3 MODEL

3.1 PROBLEM STATEMENT

In this system, each individual represents an integrated agent who formulates their actions based on their individual goals, their beliefs about the possible actions of another agent, and a complete observation of the environment. The robot, on the other hand, has access to a single demonstration of a human performing the same task in the past, as well as a complete observation of the environment. Consequently, at each step, the robot receives observations from the system and issues action commands to control the avatar within the environment. Formally, each task is defined by the person's goal g (a set of goal predicates), a demonstration of Alice taking actions to achieve that goal $D = \{\mathbf{s}^t, \mathbf{a}^t\}_{t=1}^{T}$ (a sequence of states $\mathbf{s}^t$ and actions $\mathbf{a}^t$) and a new environment where the robot collaborates with the person and help achieve the same goal as quickly as possible. During training, the ground-truth goal of the person is shown to the robot as supervision; during testing, the robot no longer has access to the ground-truth goal and thus has to infer it from the given demonstration.

3.2 LOGIC-BASED MODEL

Consider a set of objects, denoted as $\mathcal{C}$. For the object $c \in \mathcal{C}$, its trajectories and the key actions observed up to t can be summarized as a sequence of temporally ordered events $\mathcal{H}_t^c$, where $t \in R^+$ is the time. The spatial-temporal predicates include spatial-temporal *property* predicates and spatial-temporal *relation* predicates. Specifically, the spatial-temporal *property* predicates are defined as $X(\cdot)$, which illustrates some intrinsic properties of objects, such as "move slowly". The spatial-temporal *relation* predicates ($e.g.$ "book" is inside the "bookshelf") are introduced to define the spatial and temporal relations of two entities. It is noteworthy that all these boolean predicates can be converted to probabilistic ones. We can soften these logic functions by kernel functions with learnable parameters to tolerate uncertainties in data.

In general, the logic rule in our paper is defined as a logical connective of predicates, including property predicates and spatial-temporal relation predicates. For example, a sensible rule will look like:

$$\begin{aligned} Watch(person, TV) \leftarrow & Walk_to(person, bedroom) \bigwedge Switch_on(person, TV) \\ & \bigwedge Sit_on(person, sofa) \bigwedge Move_slowly(person) \end{aligned} \tag{1}$$

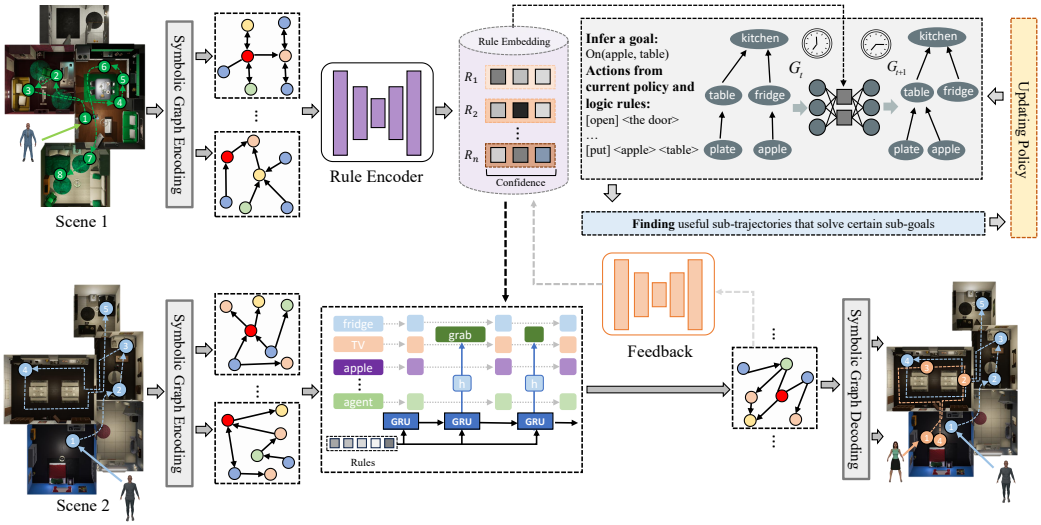


Figure 2: The overview of our proposed framework. In the first stage, given the historical motion trajectory of a person on a scene, we first utilize the symbolic graph to represent objects as nodes and the spatial relationship between them as edges. The rule encoder generates logic rules with their scores to infer the goal iteratively that the person was trying to achieve. In the second stage, the AI agent is asked to work together with the person to achieve the same goal in a new environment. The new scene is also encoded as graphs, which will pass through the propagation module to generate the AI agent's actions based on its latest observation and logic rules.

which means that a person walked slowly into the bedroom, switched on the television, and sat on the sofa. From this, we can infer that the person intends to watch TV. This logical rule may seem simple, but it applies to various different scenarios because most people, when wanting to watch TV, go through these stages.

In our study, we aim to develop a model that is informed by logical reasoning, with the capability to predict and rationalize the agent's actions. This model utilizes a query vector $\mathbf{v} = (a, s)$, comprising the action a and the state s, along with the historical trajectory τ_t observed up to the current timestep t. The integration of these elements enables our model to provide insightful explanations for the agent's behavior within the given context.

The main idea of logic-informed feature is to construct the model features using logic rules. Intuitively, given the entire trajectories τ_t and the query $\mathbf{v} = (a, s)$, the body part of the rule defines the evidence to be selectively gathered from *history* to deduce the next state ($e.g.$ open, switch-on) for query entity. Assume that for each possible symbolic state $g \in \mathcal{G}$, where g serves as the *head predicate* and $\mathcal{G}$ represents *the set of head predicates*.

Given an individual rule, we propose to build the *feature* that is conditional on history and query as $\xi(g \mid \mathbf{v}, \tau_t)$. Suppose there is a rule set $\mathcal{F}_g$, where the event g is the head predicate (goal). All the rules will play together to reason about the occurrence of g. For each $f \in \mathcal{F}_g$, one can compute the features as above. Given the rule set $\mathcal{F}_g$, we model the probability of the event g as a log-linear function of the features: $p_\alpha(g \mid \mathbf{v}, \tau_t) \propto \exp(\sum_{f \in \mathcal{F}_g} \alpha_f \cdot \xi_f(g \mid \mathbf{v}, \tau_t))$, where $\alpha = [\alpha_f]_{f \in \mathcal{F}} \geq 0$ are the learnable weight parameters associated with each rule. All the model parameters can be learned by maximizing the likelihood. We optimize the logic-based model by maximizing the likelihood of data as $\mathbb{E}_{(g,\mathbf{v},\tau_t)} \left[\log \mathbb{E}_{p_\theta} \left[p_\alpha \left(g \mid \mathbf{v}, \tau_t\right)\right]\right]$, where θ and α are learning weights. The detailed learning process and graph propagation module Liao et al. (2019) can be found in the Appendix. Then, we can use these interpretable logic rules to reason the actions of the human and the agent in the ToM collaboration framework.

3.3 Hierachycal ToM collaboration framework

Iterative Goal Inference By leveraging the trajectory information up to the current moment, we can derive the probability distribution for the next potential predicate state, expressed as:

$$g^* = \arg\max_g p_\alpha(g|s^t, a_H^t, \tau^t) = \arg\max_g \frac{\exp\left(\sum_{f\in\mathcal{F}_g} \alpha_f \cdot \xi_f(g \mid s^t, a_H^t, \tau^t)\right)}{\sum_{g'} \exp\left(\sum_{f\in\mathcal{F}_{g'}} \alpha_f \cdot \xi_f(g' \mid s^t, a_H^t, \tau^t)\right)}. \quad (2)$$

Here, g^* denotes the most probable goal inferred by the robot, which both the robot and the human strive to achieve within the next time horizon. Additionally, a_H^t signifies the human action at time t. $\mathcal{F}_g$ represents all possible rules in the rule set, and $\sum_{g'} \exp\left(\sum_{f\in\mathcal{F}_{g'}} \alpha_f \cdot \xi_f(\cdot)\right)$ is the partition function used to aggregate the probabilities of paths leading to the same goal.

The primary objective of this collaborative framework between the human and the robot is to identify the policy function π for the robot and assist people in accomplishing tasks within their capabilities. Additionally, the robot is required to consider the human's cognitive processes and construct a mental model of the human. This model enables the robot to anticipate the human's policy and enhance collaboration through deeper understanding and consideration.

Policy Estimation To formulate the policies for both the robot agent and the human agent model in the robot's mind, we employ the Boltzmann rationality model, where β serves as the rationality coefficient for the agent. This coefficient quantifies the extent to which the agent's choices are focused around the optimal decisions. As β approaches infinity, the agent behaves in a manner akin to a perfectly rational agent, meticulously optimizing its choices. Conversely, as β approaches 0, the agent becomes increasingly indifferent to the quality of Q. The collaboration is purely decentralized, with all agents acting based solely on their local observations and without any direct communication, except for ToM inferences from trajectory information. So the policy function for the agent i can be expressed as follows:

$$\pi_{g,i}(a \mid s^t, \omega^t, \tau^t) = \frac{\exp(\beta \sum_{\omega\sim\Omega} b^t(\omega_{-i} \mid \tau^t) Q_g(a, s^t))}{\sum_{a'\in A} \exp(\beta \sum_{\omega\sim\Omega} b^t(\omega_{-i} \mid \tau^t) Q_g(a', s^t))}. \quad (3)$$

Here, $b^t(\omega_{-i} \mid \tau^t)$ represents the agent's belief, a probability distribution used to estimate the private state ω_{-i} of the other agent based on this agent's trajectory. $Q_g(a, s^t)$ is the Q-function, dependent on the state and action, and can be represented as:

$$Q_g(a, s^t) = \sum\nolimits_{f\in\mathcal{F}_g} \alpha_f \cdot \xi_f(g \mid a, s^t, \tau_t). \quad (4)$$

It is worth noting that our policy function and Q function are all based on the current goal g in the rule $f_g \in \mathcal{F}$. In fact, one policy and Q-value for each subtask in the hierarchy: $\boldsymbol{\pi} = \{\pi_{g_0}, \ldots, \pi_{g_m}\}$ and $\boldsymbol{Q} = \{Q_{g_0}, \ldots, Q_{g_m}\}$. In each subtask (one rule in the rule set), the robot and the human need to take a set of primitive actions to reach the terminal states specified by this rule. Then, the next goal can be inferred and is conditioned on the current history and the next primitive action that either the human or the robot is expected to take. Our approach employs a hierarchical policy execution strategy, akin to standard programming languages that use a stack discipline. Each subtask policy receives a state as input and returns the name of a subtask to invoke. When a subtask is invoked, its name is pushed onto the Task-Stack, and its policy is executed until it reaches one of its terminal states. Once a subtask terminates, its name is popped off the Task-Stack. Importantly, if any subtask currently on the Task-Stack terminates, all subtasks below it are immediately aborted. Control then returns to the subtask that initially invoked the terminated subtask. This hierarchical approach enables the dynamic execution of subtasks and completes the final goal in a manner like online inference.

Belief Update However, in a ToM collaboration framework, the Q-function of each agent model should include two additional arguments to account for both agents' intentions and the estimated belief of the counterpart. The intentions of these two agents are considered as prior knowledge and are drawn from $\boldsymbol{\Omega} = \{\Omega_R, \Omega_H\}$ respectively, where Ω_R represents the robot's intention, and Ω_H represents the human's intention in the robot's mind. Besides, we employ the obverter technique (Choi et al., 2018) to enable both agent models to hold a belief $\hat{b}_{-i}$ as an estimation of the other agent's belief about their own intention ω. To model this nested belief, we do not need to use a distribution over the original intention distribution. This is because the belief update process follows

Algorithm 1 Adaptive Theory of Mind Collaboration Hierarchical Learning Algorithm

Randomly initialized network parameters $\{\boldsymbol{\psi_{(i,Q)}}, \boldsymbol{\psi_{(i,\pi)}}, \boldsymbol{\psi_{(i,\mathcal{E})}}\}$, $\{\hat{\boldsymbol{\psi}}_{(i,Q)}, \hat{\boldsymbol{\psi}}_{(i,\pi)}, \hat{\boldsymbol{\psi}}_{(1,\mathcal{E})}\}$, learning rate $\{\eta_Q, \eta_\pi, \eta_\mathcal{E}\}$

for each iteration **do**
  for i in $\{-1, 1\}$ **do**
    for episode $= 1, \ldots, m$ (where m is the number of sub-goals) **do**
      repeat
        Sample sub-goal: $g \leftarrow p\left(g \mid s^t, a^t, \tau^t\right)$
        Sample an action $\hat{a}_i^t \leftarrow \hat{\pi}_i\left(a \mid s^t, \omega^t, \tau^t\right)$
        Sample an action $\hat{a}_{-i}^t \leftarrow \hat{\pi}_{-i}\left(a \mid s^t, \omega^t, \tau^t\right)$
        Update two agent models' beliefs as Eq. (5)
      until Satisfy the sub-goal
    end for
    $y_i^{t,(k)} = r_i^{t,(k)} + \gamma \max_{a \in A_i} \hat{Q}_{\hat{\psi}_{(i,Q)}}(a, s^{t+1,(k)}, \omega^{(k)}, \hat{b}_i^{t,(k)})$ (k is the index of trajectory)
    $L^Q = \sum_t ||Q_{\psi_{(i,Q)}}(a_t^{(k)}, s_t^{(k)}, \omega^{(k)}, \hat{b}_t^{(k)}) - y_t^{(k)}||^2$ Compute Q-loss
    $L^\pi = \sum_{t,(k)} H(\hat{\pi}_{(-i,g)}(a_{-i}^t | s^{t,(k)}, \omega_{-i}^{t,(k)}, \tau^{t,(k)}), a_{-i}^{t,(k)})$ Compute policy loss
    $L^\mathcal{E} = \mathrm{KL}(\bar{b}_{-i}^{t,(k)} || \mathcal{E}_{(-i,g)}(\hat{b}_{-i}^{t-1,(k)}, a_i^{t,(k)}, s^{t,(k)}))$ Compute belief loss

$$\{\boldsymbol{\psi}_{(i,Q)}, \boldsymbol{\psi}_{(i,\pi)}, \boldsymbol{\psi}_{(i,\mathcal{E})}\} \leftarrow \{\boldsymbol{\psi}_{(i,Q)}, \boldsymbol{\psi}_{(i,\pi)}, \boldsymbol{\psi}_{(i,\mathcal{E})}\} - \nabla_{\{\boldsymbol{\psi}_{(i,Q)}, \boldsymbol{\psi}_{(i,\pi)}, \boldsymbol{\psi}_{(i,\mathcal{E})}\}} \{\eta_Q L^Q, \eta_\pi L^\pi, \eta_\mathcal{E} L^\mathcal{E}\}$$

    Update $\{\hat{\boldsymbol{\psi}}_{(i,Q)}, \hat{\boldsymbol{\psi}}_{(i,\pi)}, \hat{\boldsymbol{\psi}}_{(i,\mathcal{E})}\} \leftarrow \{\boldsymbol{\psi}_{(i,Q)}, \boldsymbol{\psi}_{(i,\pi)}, \boldsymbol{\psi}_{(i,\mathcal{E})}\}$ for Q-learning
  end for
end for

a deterministic path for rational agents when employing Bayesian inference. At each time step, the update process can be expressed as:

$$b^t(\omega_{-i} \mid \tau^t) = P(\omega_{-i} \mid \tau^{t-1}, a^t) \propto \hat{\pi}(a_{-i}^t \mid s^t, \omega_{-i}) b^{t-1}(\omega_{-i}) \tag{5}$$

In this equation, $\hat{\pi}$ represents the agent's estimation of the other agent's policy, which is learned during centralized training. Our belief update function incorporates counterfactual reasoning. Essentially, the agent explores all possible private states ω_{-i} and estimates the likelihood of the actions it observed under the assumption that a specific set of private agent states is correct. This approach involves fixing one agent model while training the other, thus avoiding non-stationarity issues that could arise from simultaneous agent updates. Consequently, the other agent model can be considered as part of the environment, ensuring the convergence of the Q-function of the learning agent. The whole algorithm is presented in Algorithm 1, where $\psi_{(i,Q)}, \psi_{(i,\pi)}, \psi_{(i,\mathcal{E})}$ represent the parameters of Q-function, policy function, and belief-updated function $\mathcal{E}$, respectively.

Training In Algorithm 1, Deep Q-learning (Mnih et al., 2013) is employed to train the Q-function Q_i. Simultaneously, supervised learning is used to train both the policy function $\hat{\pi}_{-i}$ and the belief update function $\mathcal{E}_{-i}$. The variable $\bar{b}_{-i}$, a result of discretizing b_{-i}, is provided as supervisory input to agent i streamlining the training of the belief update function $\mathcal{E}_{-i}$. Three loss terms are optimized sequentially. L^Q quantifies the disparity between Q-values, while the second term, L^π, corresponds to the softmax cross-entropy loss. The third term, $L^\mathcal{E}$, measures the Kullback-Leibler (KL) divergence used to assess the dissimilarity between the predicted partner's belief and the supervised partners' belief $\bar{b}_{-i}$.

4 EXPERIMENT

In this section, we provide some implementation details and show ablation studies to evaluate the performance of our framework on the Watch-and-help dataset (Puig et al., 2020) and HandMeThat dataset (Wan et al., 2022). We compare our model with several state-of-the-art approaches in human-AI interaction, including WAH (Puig et al., 2020), GEM (Netanyahu et al., 2022), DRRN (Hausknecht et al., 2020), Seq2Seq (Sutskever et al., 2014) and ReQueST (Reddy et al., 2020).

Table 1: Comparative performance of various rule encoder backbones on the Watch-and-Help Dataset: An evaluation across four difficulty tiers using three metrics—Average Number of Moves (AN, lower is better) for successful episodes, Success Rate (SR, higher is better); and Speedup (SU, higher is better).

Methods		Partially Observable			
		Level 1	Level 2	Level 3	Level 4
WAH	AN ↓	15.28±0.06	22.24±0.15	28.88±0.15	51.20±0.12
	SR ↑	78.61±0.02	71.02±0.08	60.13±0.11	51.48±0.04
	SU ↑	0.17±0.012	0.14±0.019	0.13±0.016	0.08±0.016
GEM	AN ↓	14.08±0.04	20.40±0.10	28.71±0.05	42.12±0.04
	SR ↑	79.27±0.02	71.32±0.04	60.99±0.07	53.97±0.06
	SU ↑	0.24±0.017	0.22±0.016	0.19±0.012	0.15±0.015
DRRN	AN ↓	13.47±0.13	20.10±0.14	28.68±0.15	38.25±0.01
	SR ↑	80.86±0.09	71.74±0.06	66.27±0.13	57.53±0.18
	SU ↑	0.25±0.015	0.19±0.017	0.17±0.004	0.16±0.009
Seq2Seq	AN ↓	12.76±0.06	19.92±0.06	28.32±0.02	34.16±0.04
	SR ↑	81.12±0.01	72.45±0.18	66.69±0.18	57.86±0.09
	SU ↑	0.38±0.003	0.28±0.008	0.26±0.007	0.21±0.014
ReQueST	AN ↓	12.68±0.03	18.72±0.10	28.12±0.12	32.92±0.18
	SR ↑	81.12±0.05	77.44±0.17	68.52±0.04	55.34±0.07
	SU ↑	0.38±0.019	0.30±0.015	0.25±0.009	0.22±0.011
Ours	AN ↓	11.40±0.04	16.21±0.15	26.16±0.02	31.10±0.05
	SR ↑	82.43±0.17	78.14±0.03	69.84±0.17	58.91±0.16
	SU ↑	0.48±0.015	0.44±0.009	0.36±0.013	0.33±0.017

For the Watch-and-help (Puig et al., 2020) dataset, we utilize success rate, average number of movements in successful episodes, Speedup, and cumulative reward as our metrics. Note that Speedup means we compare the episode length when the human and AI agent are working together (L_{help}) with the episode length when the human is working alone (L_{human}). For the HandMeThat dataset, we utilize the average score, the success rate(%) that the model achieves the goal within 40 steps, and the average number of moves to evaluate the performance.

4.1 Quantitative Analysis

We compare our method with several state-of-the-art approaches, and Table 1 presents the qualitative results in the Help stage of the Watch-and-help dataset. Note that WAH (Puig et al., 2020) represents a hybrid model of RL and planning, where an RL policy serves as the high-level policy and an RP is deployed to generate plans for each subgoal sampled from the RL-based high-level policy; DRRN (Hausknecht et al., 2020) is a choice-based text-game agent based on Deep Reinforcement Relevance Network. We observe that our method significantly outperforms all baselines measured by success rate with the fewest number of moves by using the true goals and reasoning about human's future plan, avoiding redundant actions and collisions with him. Especially, it achieves 57.91±0.16 and 68.84±0.17 success rate (%) in Level 4 and Level 3 respectively, which exceeds the second-best performance of Seq2Seq by 1.05 and 3.15. Compared with WAH, we also obtained 7.63% and 9.71% success rate improvements. This is because the rule encoder and evaluator can collaborate with each other to reduce search space and learn better rules. Figure 3 illustrates an example of collaboration. Our framework can actually predict users' intentions and represent some help strategies to achieve the same goal.

For the HandMeThat dataset, the performance of learning models in the partially-observable setting is shown in Table 2. Note that the Random model presents an agent that randomly selects a valid action at each time step. Obviously, our model performs decently well on all levels in partially-observable settings. In level 3, the inference of speaker intention gets more difficult. As a result, the performance drops significantly, but still outperforms other baselines. The average number of

Figure 3: The arrows indicate moving directions and the circles with black borders indicate moments when agents interacted with objects. When working alone (left), Alice had to search different rooms; but with Bob's help (right), Alice could finish the task much faster.

Table 2: Experiment results on HandMeThat dataset. Each model is evaluated on 4 hardness levels with 3 metrics: the Average Score (AS, higher is better), the Success Rate (SR, higher is better), and the Average Number of Moves (AN, lower is better) in successful episodes.

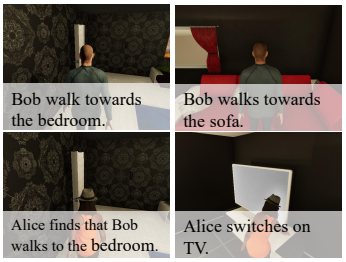
Methods		Partially Observable			
		Level 1	Level 2	Level 3	Level 4
Random	AS ↑	-40.0	-40.0	-40.0	-40.0
	SR ↑	0.0	0.0	0.0	0.0
	AN ↓	N/A	N/A	N/A	N/A
Seq2Seq	AS ↑	-5.4±0.07	-25.8±0.09	-34.7±0.15	-32.8±0.20
	SR ↑	25.50±0.20	10.40±0.14	3.95±0.23	5.30±0.26
	AN ↓	4.21±0.01	4.17±0.04	4.19±0.02	4.12±0.06
GEM	AS ↑	-6.3±0.11	-22.5±0.12	-30.9±0.14	-29.6±0.18
	SR ↑	24.64±0.40	12.47±0.21	5.74±0.15	7.21±0.13
	AN ↓	4.34±0.04	4.36±0.02	4.24±0.00	4.22±0.00
DRRN	AS ↑	-40.0	-40.0	-40.0	-40.0
	SR ↑	0.0	0.0	0.0	0.0
	AN ↓	N/A	N/A	N/A	N/A
Ours	AS ↑	-1.4±0.03	-5.8±0.06	-9.1±0.17	-11.9±0.23
	SR ↑	27.73±0.29	24.81±0.36	20.97±0.12	21.66±0.18
	AN ↓	4.05±0.01	4.09±0.01	4.14±0.03	4.21±0.02

moves when Seq2Seq succeeds is low and stable. It is possibly because the model has learned a fixed template for many pick-and-place tasks that can be accomplished in 4 or 5 steps. The reinforcement learning-based DRRN model, by contrast, totally fails in both settings, yielding a similar performance as the Random agent. This is primarily due to the large action space for the model. At each step, there are 15 to 30 valid actions, and the agent gets a sparse reward only when it successfully accomplishes the task. Therefore, in practice, we see that the random exploration strategy in the beginning stage of DRRN's training struggles to find positive rewards.

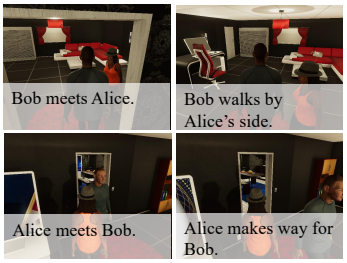
4.2 VISUALIZATION

We show some visual representations of human-robot collaboration on the Watch-and-help dataset in Figure 4. In this picture, Bob's goal is to put the plate into the microwave, and then watch TV in the bedroom. When he walks towards the plate and grabs it, Alice (human-like agent) predicts his goal as "Inside(plate, microwave)", then walks towards the microwave and opens it to help achieve the same goal as quickly as possible, so that Bob can directly put the plate into the microwave. Next, Alice sees that Bob is moving into the bedroom and getting close to the sofa, so Alice infers his goal as "Watch(TV)" based on generated logical rules. Then, Alice walks towards the TV and switches on it, so that Bob can sit down immediately and finish his goal without any redundant actions. Our rule can actually predict users' intentions and represent some help strategies to achieve the same goal with the minimum number of time steps.

Rule: Walk_to(Bob, plate)∧Grab(Bob, plate) ∧ Walk_to(Bob, microwave)∧ Open(Bob, microwave) →Inside(plate,microwave)
Explanation:
- Bob walks towards the plate and grabs it
- Bob walks towards the microwave
- Alice infers his goal as "put the plate into the microwave"
- Alice opens the microwave



Rule: Walk_to(Bob, bedroom)∧Switch_on(Alice, TV)∧ Walk_to (Bob, sofa) →Watch(Bob, TV)
Explanation:
- Bob walks towards the bedroom
- Bob walks towards the sofa
- Alice infers his goal as "watch TV"
- Alice switches on TV



Rule: Walk_to(Bob, sofa)∧Block(Alice,Bob)∧Make_way (Alice,Bob) →Pass_by(Bob,Alice)
Explanation:
- Bob walks towards the sofa
- Alice blocks Bob's way by chance
- Alice makes way for Bob so he can walk quickly
- Bob walks by Alice's side

Figure 4: Visualization and explanation of logic rules in Watch-and-help dataset.

5 LIMITATION

The present work has a few limitations. First, for real-world applications, we need to improve proposal sampling so that it can infer the graph structure and the equivalence mapping assignment more efficiently for a large number of objects. We can potentially achieve this by using language instructions to guide the inference. For instance, a human user can provide a language description of the task in addition to the physical demonstration to help achieve a better understanding of the task. Another limitation is that currently we only consider chain rules provided as prior knowledge. In the future, we plan to explore more efficient and effective rule reasoning algorithms and consider more complex rules.

6 CONCLUSION

In this paper, we introduced a novel logic-guided reasoning framework aimed at advancing human-AI collaboration. By infusing logical reasoning into AI agents, our model demonstrates a unique capability to infer and adapt to the goals of others based on observed behaviors. This is achieved by treating logic rules as latent variables and iteratively training the AI agent. In the experiments, our method can analyze the movement sequence of agents in common household activities and obtain novel insights from generated logic rules. Looking ahead, we aim to integrate additional physical laws into our framework, like the principles of energy and momentum conservation. Furthermore, we plan to introduce a predicate generator into our system. This addition will endow AI agents with advanced capabilities, enabling them to generate and reason with new predicates, thereby significantly enriching their understanding and interaction with complex environments.

REFERENCES

Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 961–971, 2016.

Chengzhi Cao, Chao Yang, and Shuang Li. Discovering intrinsic spatial-temporal logic rules to explain human actions. *arXiv preprint arXiv:2306.12244*, 2023.

Rohan Chandra, Aniket Bera, and Dinesh Manocha. Stylepredict: Machine theory of mind for human driver behavior from trajectories. *arXiv preprint arXiv:2011.04816*, 2020.

Boyuan Chen, Carl Vondrick, and Hod Lipson. Visual behavior modelling for robotic theory of mind. *Scientific Reports*, 11(1):424, 2021.

Edward Choi, Angeliki Lazaridou, and Nando de Freitas. Compositional obverter communication learning from raw visual input. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rknt2Be0-.

Wongun Choi and Silvio Savarese. Understanding collective activitiesof people from videos. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1242–1257, 2013.

Kerstin Dautenhahn. Socially intelligent robots: Dimensions of human–robot interaction. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 362(1480): 679–704, 2007. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2006.2004. URL https://royalsocietypublishing.org/doi/10.1098/rstb.2006.2004.

Artur d'Avila Garcez and Luis C. Lamb. Neurosymbolic ai: The 3rd wave, 2020.

Xuguang Duan, Xin Wang, Peilin Zhao, Guangyao Shen, and Wenwu Zhu. Deeplogic: Joint learning of neural perception and logical reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4321–4334, 2022.

Richard Evans and Edward Grefenstette. Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, 61:1–64, 2018.

Andrew Fuchs, Andrea Passarella, and Marco Conti. Modeling, replicating, and predicting human behavior: A survey. *ACM Transactions on Autonomous and Adaptive Systems*, 2023.

Riccardo Gervasi, Luca Mastrogiacomo, and Fiorenzo Franceschini. A conceptual framework to evaluate human-robot collaboration. *The International Journal of Advanced Manufacturing Technology*, 108(3):841–865, 2020. ISSN 0268-3768, 1433-3015. doi: 10.1007/s00170-020-05363-1. URL https://link.springer.com/10.1007/s00170-020-05363-1.

Michael A. Goodrich and Alan C. Schultz. Human-Robot Interaction: A Survey. *FNT in Human-Computer Interaction*, 1(3):203–275, 2007. ISSN 1551-3955, 1551-3963. doi: 10.1561/1100000005. URL http://www.nowpublishers.com/article/Details/HCI-005.

Fabio Grassiotto and Paula Dornhofer Paro Costa. Cogtom: A cognitive architecture implementation of the theory of mind. In *ICAART (2)*, pp. 546–553, 2021.

Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2255–2264, 2018.

Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Côté, and Xingdi Yuan. Interactive fiction games: A colossal adventure, 2020.

Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1971–1980, 2016.

Julian Jara-Ettinger. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29:105–110, 2019.

Henry Kautz. The third ai summer: Aaai robert s. engelmore memorial lecture. *AI Magazine*, 43(1):105–125, Mar. 2022. doi: 10.1002/aaai.12036. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/19122.

Yuan-Hong Liao, Xavier Puig, Marko Boben, Antonio Torralba, and Sanja Fidler. Synthesizing environment-aware activities via activity sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6291–6299, 2019.

Yuanyuan Mao, Shuang Liu, Pengshuai Zhao, Qin Ni, Xin Lin, and Liang He. A review on machine theory of mind. *arXiv preprint arXiv:2303.11594*, 2023.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013.

Aviv Netanyahu, Tianmin Shu, Joshua Tenenbaum, and Pulkit Agrawal. Discovering generalizable spatial goal representations via graph-based active reward learning. In *International Conference on Machine Learning*, pp. 16480–16495. PMLR, 2022.

Dung Nguyen, Phuoc Nguyen, Hung Le, Kien Do, Svetha Venkatesh, and Truyen Tran. Learning theory of mind via dynamic traits attribution. *arXiv preprint arXiv:2204.09047*, 2022.

Stefanos Nikolaidis, Ramya Ramakrishnan, Keren Gu, and Julie Shah. Efficient model learning from joint-action demonstrations for human-robot collaborative tasks. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, March 2015. doi: 10.1145/2696454.2696455. URL http://dx.doi.org/10.1145/2696454.2696455.

Uchenna Emeoha Ogenyi, Jinguo Liu, Chenguang Yang, Zhaojie Ju, and Honghai Liu. Physical human–robot collaboration: Robotic systems, learning methods, collaborative strategies, sensors, and actuators. *IEEE Transactions on Cybernetics*, 51(4):1888–1901, 2021. doi: 10.1109/TCYB.2019.2947532.

Maithili Patel and Sonia Chernova. Proactive robot assistance via spatio-temporal object modeling. *arXiv preprint arXiv:2211.15501*, 2022.

Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B Tenenbaum, Sanja Fidler, and Antonio Torralba. Watch-and-help: A challenge for social perception and human-ai collaboration. *arXiv preprint arXiv:2010.09890*, 2020.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International conference on machine learning*, pp. 4218–4227. PMLR, 2018.

Siddharth Reddy, Anca Dragan, Sergey Levine, Shane Legg, and Jan Leike. Learning human objectives by evaluating hypothetical behavior. In *International Conference on Machine Learning*, pp. 8020–8029. PMLR, 2020.

Leonel Rozo, Sylvain Calinon, Darwin G. Caldwell, Pablo Jiménez, and Carme Torras. Learning physical collaborative robot behaviors from human demonstrations. *IEEE Transactions on Robotics*, 32(3):513–527, 2016. doi: 10.1109/TRO.2016.2540623.

Tianmin Shu, Dan Xie, Brandon Rothrock, Sinisa Todorovic, and Song Chun Zhu. Joint inference of groups, events and human roles in aerial videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4576–4584, 2015.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Xiaojuan Tang, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. Rule: Neural-symbolic knowledge graph reasoning with rule embedding. *arXiv preprint arXiv:2210.14905*, 2022.

Yanming Wan, Jiayuan Mao, and Josh Tenenbaum. Handmethat: Human-robot communication in physical and social environments. *Advances in Neural Information Processing Systems*, 35:12014–12026, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/4eb33c53ed5b14ce9028309431f565cc-Paper-Datasets_and_Benchmarks.pdf.

Po-Wei Wang, Priya Donti, Bryan Wilder, and Zico Kolter. Satnet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver. In *International Conference on Machine Learning*, pp. 6545–6554. PMLR, 2019a.

Po-Wei Wang, Daria Stepanova, Csaba Domokos, and J Zico Kolter. Differentiable learning of numerical rules in knowledge graphs. In *International Conference on Learning Representations*, 2019b.

Ping Wei, Yang Liu, Tianmin Shu, Nanning Zheng, and Song-Chun Zhu. Where and why are they looking? jointly inferring human attention and intentions in complex tasks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6801–6809, 2018.

Fan Yang, Zhilin Yang, and William W Cohen. Differentiable learning of logical rules for knowledge base reasoning. *Advances in neural information processing systems*, 30, 2017.

Zhi-Hua Zhou. Abductive learning: towards bridging machine learning and logical reasoning. *Science China Information Sciences*, 62:1–3, 2019.