Joint Inference of States, Robot Knowledge, and Human (False-)Beliefs

Tao Yuan[1]   Hangxin Liu[1]   Lifeng Fan[1]   Zilong Zheng[1]   Tao Gao[1,3]   Yixin Zhu[1,2]   Song-Chun Zhu[1,2]

Abstract— Aiming to understand how human (false-)belief—a core socio-cognitive ability—would affect human interactions with robots, this paper proposes to adopt a graphical model to unify the representation of object states, robot knowledge, and human (false-)beliefs. Specifically, a parse graph (*pg*) is learned from a single-view spatiotemporal parsing by aggregating various object states along the time; such a learned representation is accumulated as the robot's knowledge. An inference algorithm is derived to fuse individual *pg* from all robots across multi-views into a joint *pg*, which affords more effective reasoning and inference capability to overcome the errors originated from a single view. In the experiments, through the joint inference over *pg*s, the system correctly recognizes human (false-)belief in various settings and achieves better cross-view accuracy on a challenging small object tracking dataset.

I. INTRODUCTION

The seminal Sally-Anne [1] study has spawned a vast research literature in developmental psychology regarding Theory of Mind (ToM); in particular, human's socio-cognition in understanding *false-belief*—the ability to understand other's belief about the world may contrast with the true reality. A cartoon version of the Sally-Anne test is shown in the left of Fig. 1: Sally puts her marble in the box and left. While Sally is out, Anne moves the marble from the box to a basket. The test would ask a human participant where Sally would look for her marble when she is back. In this experiment, the marble would still be inside the box according to Sally's false-belief, even though the marble is actually inside the basket. To answer this question correctly, an agent should understand and disentangle the object state (observation from the current frame), the (accumulated) knowledge, the belief of other agents, the ground-truth/reality of the world, and importantly, the concept of false-belief.

The prior study suggests that at the age of 4 years old, children begin to develop the capability to understand false-belief [2]. Such abilities to ascribe the mental belief to the human mind, to differentiate belief from the physical reality, and even to recognize false-belief and perform psychological reasoning, is a significant milestone in the acquisition of ToM [3], [4]. Such evidence emerged from developmental psychology in the past few decades call for integrating such socio-cognitive aspects into a modern social robot [5].

In fact, false-belief is not rare in our daily life. Two examples are depicted in the middle and the right of Fig. 1:
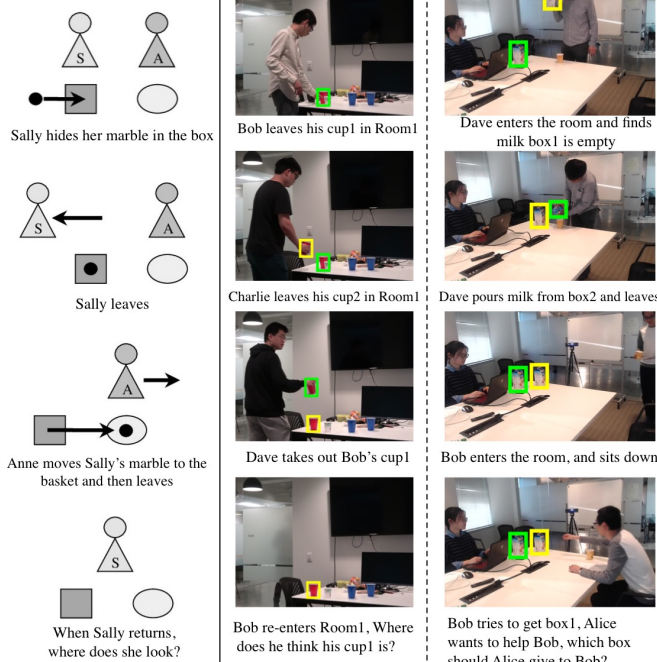
Fig. 1: Left: Illustration of the classic Sally-Anne test [1]. Middle and Right: Two different types of false-belief scenarios in our dataset: belief test and helping test.

(i) Where does Bob think his cup1 is after Charlie put cup2 (visually identical to cup1) on the table while Dave took cup1 away? (ii) Which milk box should Alice give to Bob if she wants to help? The one closer to Bob but empty, or the one further to Bob but full? Although such false-belief tasks are primal examples for social and cognitive intelligence, current state-of-the-art intelligent systems are still facing challenges in acquiring such a capability in the wild with noisy visual input (see *Related Work* for discussion).

One fundamental challenge is the lack of proper representation for modeling the false-belief from visual input; it has to be able to handle the heterogeneous information of a system's current states, its accumulated knowledge, agent's belief, and the reality/ground-truth of the world. Without a unified representation, the information across all these domains cannot be easily interpreted, and the cross-domain reasoning of the events is infeasible.

Largely due to this difficulty, prior work that takes noisy sensory input can only solve a sub-problem in understanding false-belief. For instance, sensor fusion techniques are mainly used to obtain better state estimation by filtering the measurements from multiple sensors [6]. Similarly, the Multiple View Tracking (MVT) in computer vision is designed to combine the observations across camera views to better track an object. Visual cognitive reasoning (*e.g.*, human intention/attention predictions [7], [8], [9], [10]) only

targets to model human mental states. These three lines of work are all crucial ingredients but developed independently; a unified cross-domain representation is still largely missing.

In order to endow such an ability to understand the concept of false-belief to a robot system from noisy visual inputs, this paper proposes to use a graphical model represented by a parse graph (*pg*) [11] to serve as the unified representation of a robot's knowledge structure, fused knowledge across all robots, and the (false-)beliefs of human agents. A *pg* is learned from the spatiotemporal transition of humans and objects in the scene perceived by a robot. A joint *pg* can be induced by merging and fusing the individual *pg* from each robot to overcome the errors originated from a single view. In particular, our system enables the following three capabilities with increasing depth in cognition:

1) *Tracking small objects with occlusions across different views*. Human-made objects in an indoor environment (*e.g.*, cups) are oftentimes small with a similar appearance. Tracking such objects could be challenging due to occlusions with frequent human interactions. The proposed method can address the challenging multi-view multi-object tracking problem by properly maintaining cross-view object states using the unified representation.

2) *Inferring human beliefs*. The state of an object normally does not change unless a human interacts with it; this observation shares a similar spirit in human cognition known as object permanence [12]. By identifying the interactions between humans and objects, our system also supports the high-level cognitive capability; *e.g.*, knowing which object is interacted with which person, whether a person knows the state of the object has been changed.

3) *Assisting agents by recognizing false-belief*. Giving the above object tracking and cognitive reasoning of human beliefs, the proposed algorithm can further infer whether and why the person has false-belief, thereby to better assist the person given a specific context.

A. Related Work

Robot ToM, aiming at understanding human beliefs and intents, receives increasing research attentions in human-robot interaction and collaboration [13], [14]. Several false-belief tasks akin to the classic Sally-Anne test were designed. For instance, Warnier *et al.* [15] introduced a belief management algorithm, and the reasoning capability is subsequently endowed to a robot to pass the Sally-Anne test [16] successfully. More sophisticated human-robot collaboration is achieved by maintaining a human partner's mental state [17]. More formally, Dynamic Epistemic Logic is introduced to represent and reason about belief and false-belief [18], [19]. These successes are, however, limited to the symbolic-based belief representations, requiring handcrafted variables and structures, making the logic-based reasoning approaches brittle in practice to handle noises and errors. To address this deficiency, this paper utilizes a unified representation by *pg*, a probabilistic graphical model that has been successfully applied to various robotics tasks, *e.g.*, [20], [21], [22]; it accumulates the observations over time to form a knowledge graph and robustly handles noisy visual input.

Multi-view Visual Analysis is widely applied to 3D reconstruction [23], object detection [24], [25], cross-view tracking [26], [27], and joint parsing [28]. Built on top of these modules, Multiple Object Tracking (MOT) usually utilizes tracking-by-detection techniques [29], [30], [31]. This line of work primarily focuses on combining different camera views to obtain a more comprehensive tracking, lacking the understanding of human (false-)belief.

Visual Cognitive Reasoning is an emerging field in computer vision. Related work includes recovering incomplete trajectories [32], learning utility and affordance [33], inferring human intention and attention [9], [10], *etc*. As to understanding (false-)belief, despite many psychological experiments and theoretical analysis [34], [35], [36], [37], very few attempts have been made to solve (false-)belief with visual input; handcrafted constraints are usually required for specific problems in prior work. In contrast, this paper utilizes a unified representation across different domains with heterogeneous information to model human mental states.

B. Contribution

This paper makes three contributions:

1) We adopt a unified graphical model *pg* to represent and maintain heterogeneous knowledge about object states, robot knowledge, and human beliefs.

2) On top of the unified representation, we propose an inference algorithm to merge individual *pg* from different domains across time and views into a joint *pg*, supporting human belief inference from multi-view to overcome the noises and errors originated from a single view.

3) With the inferred *pg*s, our system can keep track of the state and location of each object, infer human beliefs, and further discover false-belief to better assist human.

C. Overview

The remainder of the paper is organized as follows. Sections II and III describe the representation and the detailed probabilistic formulation, respectively. We demonstrate the efficacy of the proposed method in Section IV and conclude the paper with discussions in Section V.

II. REPRESENTATION

In this work, we use the parse graph (*pg*)—a unified graphical model [11]—to represent (i) the location of each agent and object, (ii) the interactions between agents and objects, (iii) the beliefs of agents, and (iv) the attributes and states of objects; see Fig. 2 for an example. Specifically, three different types of *pg*s are utilized:

- *Robot pg*, shown as blue circles, maintains the knowledge structure of an individual robot, which is extracted from its visual observation—an image. It also contains attributes that are grounded to the observed agents and objects.
- *Belief pg*, shown as red diamonds, represents the *inferred* human knowledge by each robot. Each robot maintains the parse graph for each agent it observed.
- *Joint pg* fuses all the information and views across a set of distributed robots.

Notations and Definitions: The input of our system can be represented by M synchronized video sequences $I = \{I_{t=1..T}^{k=1..M}\}$ with length T captured from M robots. Formally, a scene $\mathcal{R}$ is expressed as

$$\mathcal{R} = \{(O_t, H_t) : t = 1, 2, \ldots, T\},$$
$$O_t = \{o_t^i : i = 1, 2, \ldots, N_o\}, \qquad (1)$$
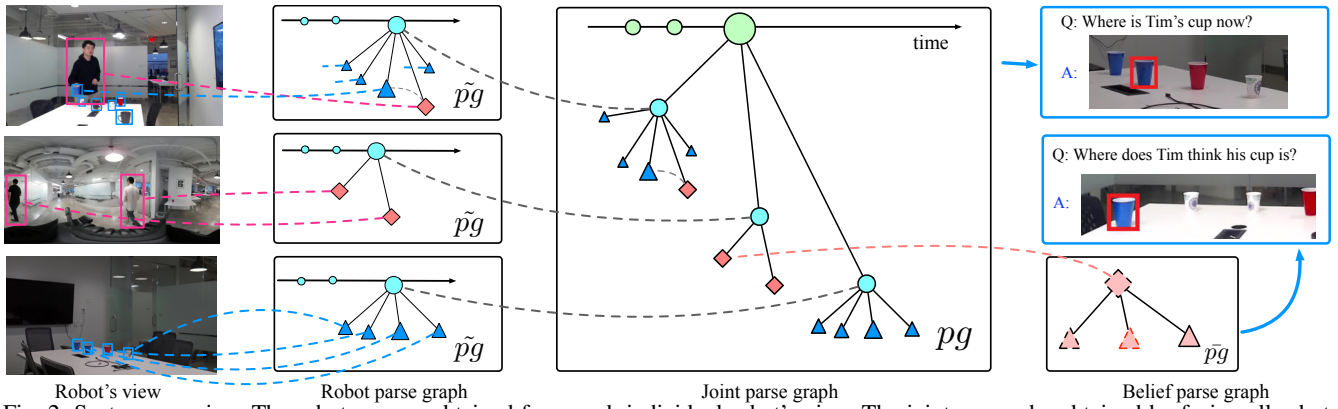$$H_t = \{h_t^j : j = 1, 2, \ldots, N_h\},$$

Fig. 2: System overview. The robot *pg*s are obtained from each individual robot's view. The joint *pg* can be obtained by fusing all robots' *pg*s. The belief *pg*s can be inferred from the joint *pg*. All the *pg*s are optimized simultaneously under the proposed joint parsing framework to enable the queries about the object states and human (false-)beliefs.

where O_t and H_t denote the set of all the tracked objects (N_o objects in total) and the set of all the tracked agents (N_h agents in total) at time t, respectively.

Object o_t^i is represented by a tuple: bounding box location b_t^i, appearance feature ϕ_t^i, states s_t^i, and attributes a_t^i,

$$o_t^i = (b_t^i, \phi_t^i, s_t^i, a_t^i), \qquad (2)$$

where s_t^i is an index function: $s_t^i = j, j \neq 0$ indicates the object o_i is held by the agent h_j at time t, and $s_t^i = 0$ means it is not held by any agent at time t.

The agent h_t^j is represented by its body key-point position κ_t^j and appearance feature ϕ_t^j

$$h_t^j = (\kappa_t^j, \phi_t^j). \qquad (3)$$

Robot Parse Graph is formally expressed as

$$\tilde{pg}_t^k = \{(o_t^i, h_t^j) : o_t^i, h_t^j \in I_t^k\}, \qquad (4)$$

where I_t^k is the area where kth robot can observe at time t.

Belief Parse Graph is formally expressed as

$$\bar{pg}_t^{k,j} = \{o_{t'}^i : o_{t'}^i \in I_{t'}^k\}, \qquad (5)$$

where $\bar{pg}_t^{k,j}$ represents the inferred belief of agent h_j under robot k's view; t' is the last time that the robot k observes the human h_j. We assume that the agent h_j only keeps the objects s/he observed last time in this area in mind, which satisfies the *Principle of Inertia*: an agent's belief is preserved over time unless the agent gets information to the contrary.

Joint Parse Graph keeps track of all the information across a set of distributed robots, formally expressed as

$$pg_t = \{(o_t^i, h_t^j : i = 1, 2, ..., N_o; j = 1, 2, ..., N_h)\}. \qquad (6)$$

Objective: The objective of the system is to jointly infer all the parse graphs $PG = \{pg, \tilde{pg}, \bar{pg}\}$ so that it can (i) track all the agents and objects across scenes at any time by fusing the information collected by a distributed system, and (ii) infer human (false-)beliefs to provide assistance.

III. PROBABILISTIC FORMULATION

We formulate the joint parsing problem as a maximizing a posterior (MAP) inference problem

$$PG^* = \arg\max_{PG} p(PG|I) = \arg\max_{PG} p(I|PG) \cdot p(PG), \qquad (7)$$

where $p(PG)$ is the prior, and $p(I|PG)$ is the likelihood.

A. Prior

The prior term $p(PG)$ models the compatibility of the robot *pg*s and the joint *pg*, and the compatibility of the joint *pg* over time. Formally, we can decompose the prior as

$$p(PG) = p(pg_1) \prod_{t=1}^{T-1} p(pg_{t+1}|pg_t) \prod_{k=1}^{M} \prod_{t=1}^{T} p(\tilde{pg}_t^k|pg_t), \qquad (8)$$

where the first term $p(pg_{t+1}|pg_t)$ is the transition probability of the joint *pg* over time, further decomposed as

$$p(pg_{t+1}|pg_t) = \frac{1}{Z} \exp\{-\mathcal{E}(pg_{t+1}|pg_t)\}, \qquad (9)$$

$$\mathcal{E}(pg_{t+1}|pg_t) = \sum_{i=1}^{N_o} \mathcal{E}_{L_o}(b_{t+1}^i, b_t^i, s_t^i) + \mathcal{E}_{ST}(s_{t+1}^i, s_t^i) + \sum_{j=1}^{N_h} \mathcal{E}_{L_h}(\kappa_{t+1}^j, \kappa_t^j). \qquad (10)$$

The second term $p(\tilde{pg}_t^k|pg_t)$ is the probability which models the compatibility of individual *pg*s and the joint *pg*. Its energy can be decomposed into three energy terms

$$p(\tilde{pg}_t^k|pg_t) = \frac{1}{Z} \exp\{-\mathcal{E}(pg_t, \tilde{pg}_t^k)\}$$

$$= \frac{1}{Z} \exp\{-\mathcal{E}_A(pg_t, \tilde{pg}_t^k) - \mathcal{E}_S(pg_t, \tilde{pg}_t^k) - \mathcal{E}_{Attr}(pg_t, \tilde{pg}_t^k)\}. \qquad (11)$$

Below, we detail the above six energy terms $\mathcal{E}_{(\cdot)}$ in Eqs. (10) and (11).

Motion Consistency: The term $\mathcal{E}_L$ measures the motion consistency of objects and agents in time, defined as

$$\mathcal{E}_{L_o}(b_{t+1}^i, b_t^i, s_t^i) = \begin{cases} \delta(\mathcal{D}(b_{t+1}^i, b_t^i) > \tau)) & \text{if } s_t^i = 0 \\ \delta(\mathcal{D}(\kappa_{t+1}^j, \kappa_t^j) > \tau)) & \text{if } s_t^i = j \end{cases} \qquad (12)$$

$$\mathcal{E}_{L_h}(\kappa_{t+1}^j, \kappa_t^j) = \delta(\mathcal{D}(\kappa_{t+1}^j, \kappa_t^j) > \tau)),$$

where $\mathcal{D}$ is the distance between two bounding boxes or human poses, τ is the speed threshold, and δ is the indicator function. If an object i is held by an agent j, we use the agent's location to calculate $\mathcal{E}_L$ of the object.

State Transition Consistency: The term $\mathcal{E}_{ST}$ is the state transition energy, defined as

$$\mathcal{E}_{ST}(s_{t+1}^i, s_t^i) = -\log p(\delta(s_{t+1}^i = 0)|\delta(s_t^i = 0)), \qquad (13)$$

where the state transition probability $p(\delta(s_{t+1}^i = 0)|\delta(s_t^i = 0))$ is learned from the training data.

Appearance Consistency: $\mathcal{E}_A$ measures appearance consistency. In robot *pg*s, the appearance feature vector ϕ

(a) Put down a cup   (b) Pick up a cup   (c) Move a cup

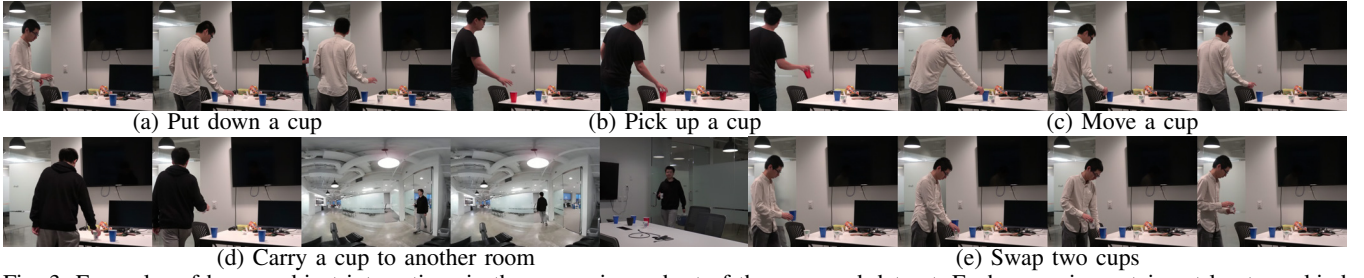(d) Carry a cup to another room   (e) Swap two cups

Fig. 3: Examples of human-object interactions in the cross-view subset of the proposed dataset. Each scenario contains at least one kind of false-belief test or helping test recorded with four robot camera views.

is extracted by a deep person re-identification network [38]. In the joint pg, the feature vector is calculated by the mean pooling of all features for the same entity from all robot pgs.

$$\mathcal{E}_A(pg_t, \tilde{pg}_t) = \sum_{e \in O_t \cup H_t} ||\phi_t^e - \phi_t^{\tilde{e}}||_2. \tag{14}$$

Spatial Consistency: Each object and agent in the robot's 2D view should also have a corresponding 3D location in the real-world coordinate system, and such correspondence should remain consistent when projected any points from the robot image plane back to the real-world coordinate. Thus, spatial consistency is defined as

$$\mathcal{E}_S(pg_t, \tilde{pg}_t) = \sum_{e \in O_t \cup H_t} ||\Lambda_t^e - f(\Lambda_t^{\tilde{e}})||_2, \tag{15}$$

where Λ is the 3D positions in the real-world coordinate, and f is the transformation function that projects the points from the robot's 2D view to the 3D real-world coordinate.

Attribute Consistency: Attributes of each entity should remain the same across time and viewpoints. Such an attribute consistency is defined by the term $\mathcal{E}_{Attr}$

$$\mathcal{E}_{Attr}(pg_t, \tilde{pg}_t) = \sum_{i=1}^{N_o} \delta(a_t^i \neq \tilde{a}_t^i). \tag{16}$$

B. Likelihood

The likelihood term $p(I|PG)$ models how well each robot can ground the knowledge in its pg to the visual data it captures. Formally, the likelihood is defined as

$$p(I|PG) = \prod_{k=1}^{M} \prod_{t=1}^{T} p(I_t^k|\tilde{pg}_t^k). \tag{17}$$

The energy of term $p(I_t^k|\tilde{pg}_t^k)$ can be further decomposed as

$$p(I_t^k|\tilde{pg}_t^k) = \frac{1}{Z} \exp\{-\mathcal{E}(I_t^k|\tilde{pg}_t^k)\}, \tag{18}$$

$$\mathcal{E}(I_t^k|\tilde{pg}_t^k) = \sum_{i=1}^{N_o} \mathcal{E}_D(b_t^i, \phi_t^i) + \mathcal{E}_C(b_t^i, \phi_t^i, a_t^i) + \sum_{j=1}^{N_h} \mathcal{E}_D(p_t^j, \phi_t^j), \tag{19}$$

where $\mathcal{E}_D$ can be calculated by the score of object detection or human pose estimation, and $\mathcal{E}_C$ can be obtained by the object attributes classification scores.

C. Inference

Given the above probabilistic formulation, we can infer the best $\{pg^*, \tilde{pg}^*\}$ by an MAP estimate. It can be solved by two steps: (i) Each robot individually processes the visual input; the output (*e.g.*, object detection, and human pose estimation) can be aggregated as the proposals for the second step. (ii) The MAP estimate can be transformed to an assignment problem given the proposals, solvable using the *Kuhn-Munkres* algorithm [39], [40] in polynomial time.

Based on Eq. (5), robot k can generate belief parse graphs $\bar{pg}^{k,j}$ for agent j after obtaining the robot graphs $\tilde{pg}^k$.

IV. EXPERIMENT

We evaluate the proposed method in two setups: cross-view object tracking and human (false-)belief understanding. The first experiment evaluates the accuracy of object localization using the proposed inference algorithms, focusing on the robot parse graphs $\tilde{pg}$ and the joint parse graph pg. The second experiment evaluates the inference of the belief parse graphs $\bar{pg}$, *i.e.*, human beliefs regarding the object states (*e.g.*, locations) in both single-view and multi-view settings.

A. Dataset

The dataset includes two subsets, a multi-view subset and a single-view subset. Ground-truth tracking results of objects and agents, and states and attributes of objects are all annotated for evaluation purpose.

- The single-view subset includes 5 different false-belief scenarios with 12532 frames. Each scenario contains at least one kind of false-belief test or helping test. In this subset, objects are not limited to the cups.
- The multi-view subset consists of 8 scenes, each shot with 4 robot camera views, making a total number of 72720 frames. Each scenario contains at least one kind of false-belief test. The objects in each scene are, however, limited to the cups: 12-16 different cups made with 3 different materials (plastic, paper, and ceramic) and 4 colors (red, blue, white, and black). In each scene, three agents interact with cups by performing actions depicted in Fig. 3.

B. Implementation Details

Below, we detail the implementations of the system.

- Object detection: we use the RetinaNet model [41] pre-trained on the MS COCO dataset [42]. We keep all the bounding boxes with a score higher than the threshold 0.2, which serve as the proposals for object detection.
- Human pose estimation: we apply the AlphaPose [43].
- Object attribute classification: A VGG16 network [44] was trained to classify the color and the material of the objects.
- Appearance feature: A deep person re-id model [38] was fine-tuned on the training set.
- Due to the lack of multi-view in the single-view setting, we locate the object that an agent plan to interact by simply finding the object closest to the direction the agent points at according to the key points on the arm.

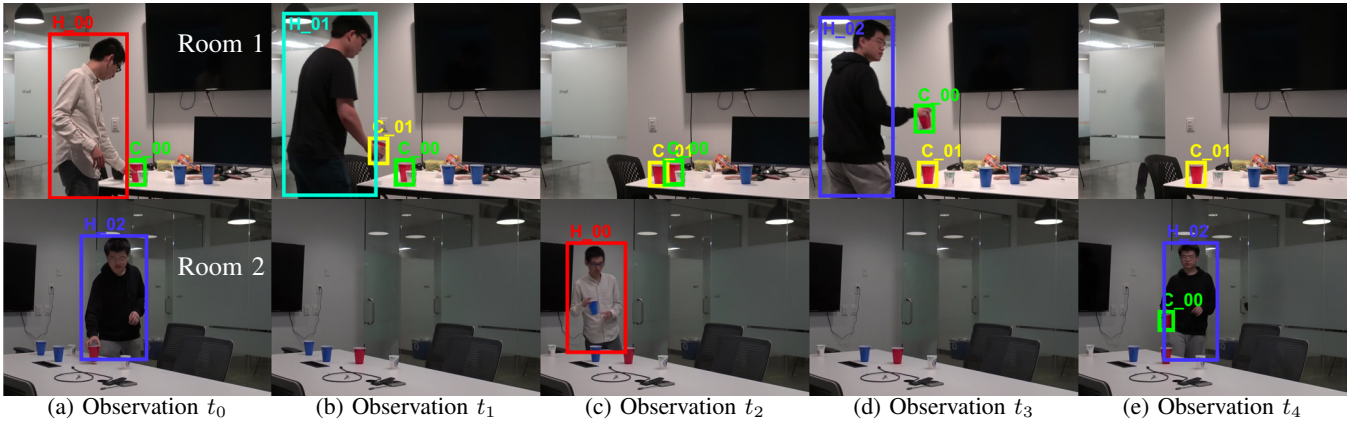| (a) Observation t_0 | (b) Observation t_1 | (c) Observation t_2 | (d) Observation t_3 | (e) Observation t_4 |

Fig. 4: A sample result in Experiment 1. Two rows show examples of synchronized visual inputs in different rooms. Two red cups marked by green and yellow bounding boxes are highlighted. At time t_0, agent H_{00} put a red cup C_{00} in Room 1. At t_1, agent H_{01} put another red cup C_{01} in Room 1. At t_3, agent H_{02} took away cup C_{00} and put it in Room 2 at t_4. Our system can robustly perform such a complex multi-room multi-view tracking and reason about agent's belief. For instance, at t_4, the tracking system knows that the cup C_{00} appeared at t_0 is in fact in Room 2 and inferred that human H_{00} thinks the cup is still in Room 1.

C. Experiment 1: Cross-view Object Localization

To test the overall cross-view tracking performance, 2000 queries are randomly sampled from the ground-truth tracks. Each query q can be formally described as

$$q = (k, t, b, t_q), \qquad (20)$$

where the tuple (k, t, b) indicates the object shown in robot k's view located in bounding box b at time t. Such a form of the query can be very flexible. For instance, if we ask about the location of that object at time t_q, the system should return an answer in the form of (k_a, b_a), meaning that the system predicts the object is shown in robot k_a's view at b_a.

The system generates the answer in two steps. It firstly locates the query of the object by searching the object i in $\tilde{p}g_t^k$ with the smallest distance to the bounding box b. Then it returns the location $b_{t_q}^i$ from $\tilde{p}g_{t_q}^{k'}$. The accuracy of model M can be calculated as

$$acc(M) = \frac{1}{N_q} \sum_{i=1}^{N_q} \delta(\text{IoU}(b_{gt}^i, b_a^i) > \xi) \cdot \delta(k_{gt} = k_a), \quad (21)$$

where N_q is the number of queries, b_{gt} is the ground-truth bounding box, and b_a is the inferred bounding box returned by model M. We calculate the Intersection over Union (IoU) between the answer and the ground-truth bounding boxes; the answer is correct if and only if the answer predicts the right view and the IoU is larger than $\xi = 0.5$.

Table I shows the ablative study by turning on and off the joint parsing component that models human interactions, *i.e.*, whether the model parses and tracks objects by reasoning about the interaction with agents. "# interactions" means how many times the object was interacted by agents. The result shows that our system achieves an overall 88% accuracy. Even without parsing humans, our system still possesses the ability to reason about object location by maintaining other consistencies, such as spatial consistency and appearance consistency. However, its performance drops significantly if the object was moved to different rooms. Figure 4 shows some qualitative results.

D. Experiment 2: (False-)Belief Inference

In this experiment, we evaluate the performance of belief and false-belief inference, *i.e.*, whether an agent's belief pg is the same as the true object states. The evaluations were conducted on both single-view and multi-view scenarios.

Multi-view: We collected 200 queries with ground-truth annotations that focus on the Sally-Anne false belief task. The query is defined as

$$q = (k_o, t_o, b_o, k_h, t_h, b_h, t_q), \qquad (22)$$

where first three terms (k_o, t_o, b_o) define the objects in robot k_o's view located at b_o at time t_o. Similarly, another three terms (k_h, t_h, b_h) define an agent in robot k_h's view located at b_h at time t_h. The question is: where does the agent (k_h, t_h, b_h) think the object (k_o, t_o, b_o) is at time t_q?

Our system generates the answer in three steps: (i) search for the object i and the agent j in robot parse graphs $\tilde{p}g_{t_o}^{k_o}$ and $\tilde{p}g_{t_h}^{k_h}$, (ii) retrieve all the belief parse graphs $\bar{p}g_{t_q}^{k',j}$ at time t_q to find the object i's location $\bar{b}_{t_q}^i$ in human j's belief, and (iii) find an object i' in robot parse graph, which has the same attributes as i's and has smallest distance to $\bar{b}_{t_q}^i$. The system finally returns i''s location $b_{t_q}^{i'}$ as the answer.

Since there is no publicly available code on this task, we compare our inference algorithm with a random baseline model as the reference for future benchmark; it simply returns an object with the same attributes as the query object at t_q. The result shows that our system achieves 81% accuracy, while the baseline model only has 39% accuracy.

Single-view: We collected a total of 100 queries, including two types of belief inference tasks: the Sally-Anne false-belief task and the helping task, as shown in Fig. 1. The queries have two forms

$$q = (t_o, b_o, b_h, t_q), \quad and \quad q = (b_h, t_q), \qquad (23)$$

indicating two different types of questions: (i) where does the agent b_h think the object (t_o, b_o) is at time t_q? (ii)

TABLE I: Accuracy of cross-view object tracking

# interactions	0	1	2	3	Overall
Parsing w/o humans acc.	0.98	0.82	0.78	0.75	0.82
Joint parsing acc.	0.98	0.86	0.85	0.82	0.88

TABLE II: Accuracy of belief queries on single view subset

	True Belief	False-Belief	Overall
Joint parsing acc.	0.94	0.93	0.94
Random guessing acc.	0.45	0.53	0.46

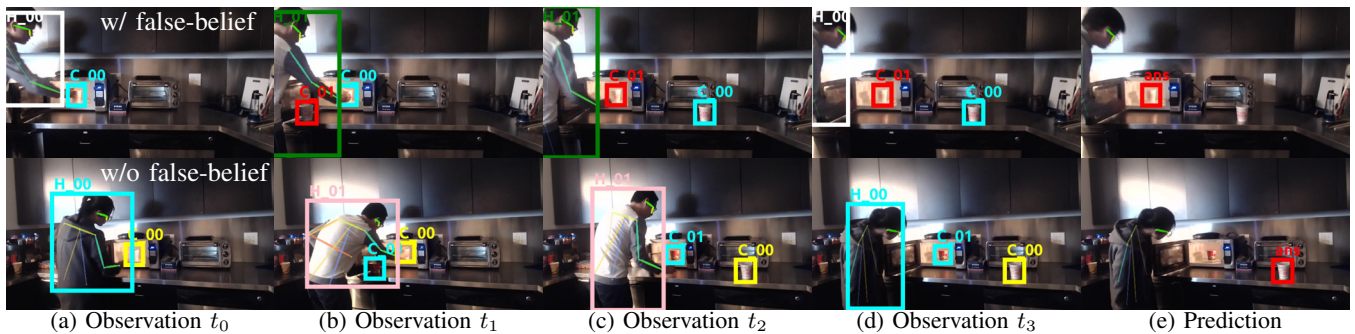(a) Observation t_0 (b) Observation t_1 (c) Observation t_2 (d) Observation t_3 (e) Prediction

Fig. 5: Two sample results of the Sally-Anne false-belief task in Experiment 2. (Top) with false-belief. (Bottom) Without false-belief. At time t_0, the first person H_{00} put the cup noodle (C_{00}) into the microwave and leaves. The second person H_{01} take out the cup noodle C_{00} from the microwave at t_1. At time t_2, H_{01} put his own cup noodle C_{01} in the microwave and left. When H_{00} returns to the room at t_3, our system is able to answer: where does the first person H_{00} think the cup noodle C_{00} is. It can successfully predict the person in the bottom row will not have the false-belief due to the different color attributes of the cups.
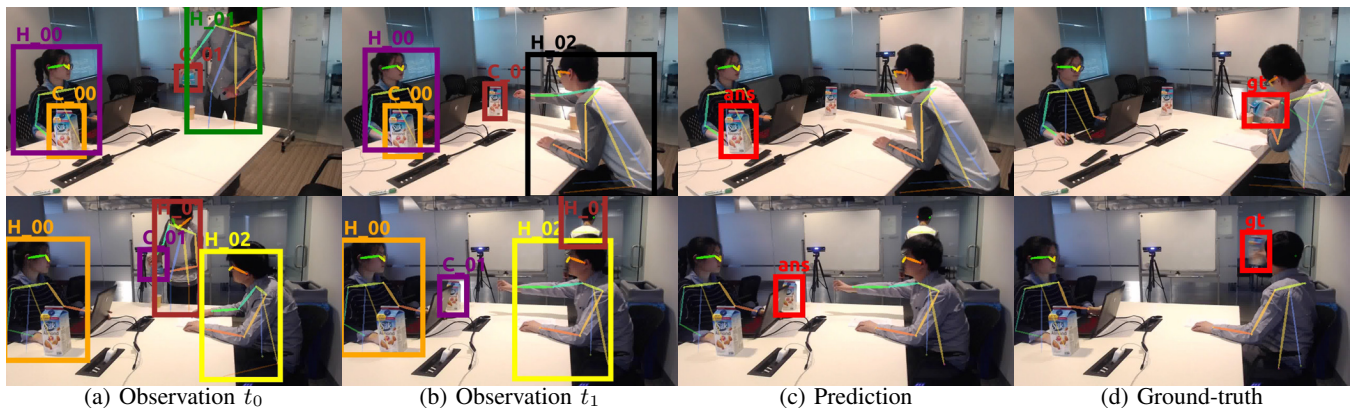

(a) Observation t_0 (b) Observation t_1 (c) Prediction   (d) Ground-truth

Fig. 6: Two sample results of the helping task in Experiment 2. (Top) At time t_0, the second person (H_{01}) enters the room and empties the box C_{01}. Question: which box should the first person (H_{00}) give to the third person (H_{02}) if the first person (H_{00}) wants to help at t_1? Answer returned by the system correctly infers that H_{02} has a false-belief, and H_{00} should give another box (C_{00}) to H_{02}, rather than the C_{01} his is reaching for. (Bottom) Since the person H_{02} observes the entire process, there is no false-belief; in this case, H_{02} reaches for C_{01} is to throw it away.

Which object will you give to the agent (t_q, b_h) at time t_q if you would like to help? For the first type of questions, i.e., the Sally-Anne false-belief task, similar to the multi-view setting, the system should return the object bounding box as the answer. For the second type of question, i.e., the helping task, the system first infers whether the agent has false-belief. If not, the system returns the object the person wants to interact based on their current pose; otherwise, the system returns another suitable object closest to them. Qualitative results are shown in Figs. 5 and 6, and quantitative results are provided in Table II.

V. CONCLUSION AND DISCUSSIONS

In this paper, we describe the idea of using *pg* as a unified representation for tracking object states, accumulating robot knowledge, and reasoning about human (false-)beliefs. Based on the spatiotemporal information observed from multiple camera views of one or more robots, robot *pg* and belief *pg* are induced and merged to a joint *pg* to overcome the possible errors originated from a single view. With such a representation, a joint inference algorithm is proposed, which possesses the capabilities of tracking small occluded objects across different views and inferring human (false-)beliefs. In experiments, we first demonstrate that the joint inference over the merged *pg* produced better tracking accuracy. We

further evaluate the inference on human true- and false-belief regarding objects' locations by jointly parsing the *pg*s. The high accuracy demonstrates that our system is capable of modeling and understanding human (false-)beliefs, with the potential of helping capability as demonstrated in developmental psychology.

ToM and Sally-Anne test are interesting and difficult problems in the area of social robotics. For a service robot to interact with humans in an intuitive manner, it must be able to maintain a model of the belief states of the agents it interacts with. We hope the proposed method using a graphical model has demonstrated a different perspective compared to prior methods in terms of flexibility and generalization. In the future, a more interactive and active set up would be more practical and compelling. For instance, by integrating activity recognition modules, our system should be able to perceive, recognize, and extract richer semantic information from the observed visual input, thereby providing more subtle (false-)belief applications. Communication, gazes, and gestures are also crucial in intention expression and perception in collaborative interactions. By incorporating these essential ingredients and taking advantage of the flexibility and generalization of the model, our system should be able to go from the current passive query to active response to assist agents in real-time.

References

[1] S. Baron-Cohen, A. M. Leslie, and U. Frith, "Does the autistic child have a theory of mind?," *Cognition*, vol. 21, no. 1, pp. 37–46, 1985.

[2] A. Gopnik and J. W. Astington, "Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction," *Child development*, pp. 26–37, 1988.

[3] O. N. Saracho, "Theory of mind: children's understanding of mental states," *Early Child Development and Care*, vol. 184, no. 6, pp. 949–961, 2014.

[4] H. Wimmer and J. Perner, "Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception," *Cognition*, vol. 13, no. 1, pp. 103–128, 1983.

[5] C. Breazeal, J. Gray, and M. Berlin, "An embodied cognition approach to mindreading skills for socially intelligent robots," *International Journal of Robotics Research*, vol. 28, no. 5, pp. 656–680, 2009.

[6] M. Liggins II, D. Hall, and J. Llinas, *Handbook of multisensor data fusion: theory and practice*. CRC press, 2017.

[7] H. Koppula and A. Saxena, "Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation," in *International Conference on Machine Learning (ICML)*, 2013.

[8] S. Qi, S. Huang, P. Wei, and S.-C. Zhu, "Predicting human activities using stochastic grammar," in *International Conference on Computer Vision (ICCV)*, 2017.

[9] L. Fan, Y. Chen, P. Wei, W. Wang, and S.-C. Zhu, "Inferring shared attention in social scene videos," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[10] P. Wei, Y. Liu, T. Shu, N. Zheng, and S.-C. Zhu, "Where and why are they looking? jointly inferring human attention and intentions in complex tasks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[11] S.-C. Zhu, D. Mumford, *et al.*, "A stochastic grammar of images," *Foundations and Trends® in Computer Graphics and Vision*, vol. 2, no. 4, pp. 259–362, 2007.

[12] R. Baillargeon, E. S. Spelke, and S. Wasserman, "Object permanence in five-month-old infants," *Cognition*, vol. 20, no. 3, pp. 191–208, 1985.

[13] B. Scassellati, "Theory of mind for a humanoid robot," *Autonomous Robots*, vol. 12, no. 1, pp. 13–24, 2002.

[14] A. Thomaz, G. Hoffman, M. Cakmak, *et al.*, "Computational human-robot interaction," *Foundations and Trends® in Robotics*, vol. 4, no. 2-3, pp. 105–223, 2016.

[15] M. Warnier, J. Guitton, S. Lemaignan, and R. Alami, "When the robot puts itself in your shoes. managing and exploiting human and robot beliefs," in *International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2012.

[16] G. Milliez, M. Warnier, A. Clodic, and R. Alami, "A framework for endowing an interactive robot with reasoning capabilities about perspective-taking and belief management," in *International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2014.

[17] S. Devin and R. Alami, "An implemented theory of mind to improve human-robot shared plans execution," in *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016.

[18] T. Bolander, "Seeing is believing: Formalising false-belief tasks in dynamic epistemic logic," in *Jaakko Hintikka on Knowledge and Game-Theoretical Semantics*, pp. 207–236, Springer, 2018.

[19] E. Lorini and F. Romero, "Decision procedures for epistemic logic exploiting belief bases," in *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2019.

[20] M. Edmonds, F. Gao, X. Xie, H. Liu, S. Qi, Y. Zhu, B. Rothrock, and S.-C. Zhu, "Feeling the force: Integrating force and pose for fluent discovery through imitation learning to open medicine bottles," in *International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[21] H. Liu, Y. Zhang, W. Si, X. Xie, Y. Zhu, and S.-C. Zhu, "Interactive robot knowledge patching using augmented reality," in *International Conference on Robotics and Automation (ICRA)*, 2018.

[22] M. Edmonds, F. Gao, H. Liu, X. Xie, S. Qi, B. Rothrock, Y. Zhu, Y. N. Wu, H. Lu, and S.-C. Zhu, "A tale of two explanations: Enhancing human trust by explaining robot behavior," *Science Robotics*, vol. 4, no. 37, 2019.

[23] M. Hofmann, D. Wolf, and G. Rigoll, "Hypergraphs for joint multi-view reconstruction and multi-object tracking," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[24] J. Liebelt and C. Schmid, "Multi-view object class detection with a 3d geometric model," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[25] A. Utasi and C. Benedek, "A 3-d marked point process model for multi-view people detection," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[26] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 33, no. 9, pp. 1806–1819, 2011.

[27] Y. Xu, X. Liu, Y. Liu, and S.-C. Zhu, "Multi-view people tracking via hierarchical trajectory composition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[28] H. Qi, Y. Xu, T. Yuan, T. Wu, and S.-C. Zhu, "Scene-centric joint parsing of cross-view videos," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[29] L. Wen, W. Li, J. Yan, Z. Lei, D. Yi, and S. Z. Li, "Multiple target tracking based on undirected hierarchical relation hypergraph," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[30] A. Dehghan, Y. Tian, P. H. Torr, and M. Shah, "Target identity-aware network flow for online multiple target tracking," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[31] X. Dong, J. Shen, D. Yu, W. Wang, J. Liu, and H. Huang, "Occlusion-aware real-time object tracking," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 763–771, 2017.

[32] W. Liang, Y. Zhu, and S.-C. Zhu, "Tracking occluded objects and recovering incomplete trajectories by reasoning about containment relations and human actions," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[33] Y. Zhu, C. Jiang, Y. Zhao, D. Terzopoulos, and S.-C. Zhu, "Inferring forces and learning human utilities from videos," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[34] J. Call and M. Tomasello, "A nonverbal false belief task: The performance of children and great apes," *Child development*, vol. 70, no. 2, pp. 381–395, 1999.

[35] C. L. Baker, R. Saxe, and J. B. Tenenbaum, "Action understanding as inverse planning," *Cognition*, vol. 113, no. 3, pp. 329–349, 2009.

[36] T. Bräuner, P. Blackburn, and I. Polyanskaya, "Second-order false-belief tasks: Analysis and formalization," in *International Workshop on Logic, Language, Information, and Computation*, 2016.

[37] Y. Wu, J. A Haque, and L. Schulz, "Children can use others' emotional expressions to infer their knowledge and predict their behaviors in classic false belief tasks," in *Annual Meeting of the Cognitive Science Society (CogSci)*, 2018.

[38] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3702–3712, 2019.

[39] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[40] H. W. Kuhn, "Variants of the hungarian method for assignment problems," *Naval Research Logistics Quarterly*, vol. 3, no. 4, pp. 253–258, 1956.

[41] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *International Conference on Computer Vision (ICCV)*, 2017.

[42] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision (ECCV)*, 2014.

[43] H. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *International Conference on Computer Vision (ICCV)*, 2017.

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.