Towards Collaborative Plan Acquisition
through Theory of Mind Modeling in Situated Dialogue

Cristian-Paul Bara[1], **Ziqiao Ma**[1], **Yingzhuo Yu**[1], **Julie Shah**[2], **Joyce Chai**[1]

[1]University of Michigan
[2]Massachussets Institute of Technology
{cpbara,marstin,yyzjason,chaijy}@umich.edu, julie_a_shah@csail.mit.com

Abstract

Collaborative tasks often begin with partial task knowledge and incomplete initial plans from each partner. To complete these tasks, agents need to engage in situated communication with their partners and coordinate their partial plans towards a complete plan to achieve a joint task goal. While such collaboration seems effortless in a human-human team, it is highly challenging for human-AI collaboration. To address this limitation, this paper takes a step towards **collaborative plan acquisition**, where humans and agents strive to learn and communicate with each other to acquire a complete plan for joint tasks. Specifically, we formulate a novel problem for agents to predict the missing task knowledge for themselves and for their partners based on rich perceptual and dialogue history. We extend a situated dialogue benchmark for symmetric collaborative tasks in a 3D blocks world and investigate computational strategies for plan acquisition. Our empirical results suggest that predicting the partner's missing knowledge is a more viable approach than predicting one's own. We show that explicit modeling of the partner's dialogue moves and mental states produces improved and more stable results than without. These results provide insight for future AI agents that can predict what knowledge their partner is missing and, therefore, can proactively communicate such information to help their partner acquire such missing knowledge toward a common understanding of joint tasks.

1 Introduction

With the simultaneous progress in AI assistants and robotics, it is reasonable to anticipate the forthcoming milestone of embodied assistants. However, a critical question arises: how can we facilitate interactions between humans and robots to be as intuitive and seamless as possible? To bridge this gap, a significant challenge lies in the mismatched knowledge and skills between humans and agents, as well as their tendency to begin with incomplete and divergent partial plans. Considering that the average user cannot be expected to possess expertise in robotics, language communications become paramount. Consequently, it becomes imperative for humans and agents to engage in effective language communication to establish shared plans for collaborative tasks. While such coordination and communication occur organically between humans, it's notoriously difficult for human-AI teams, particularly when it comes to physical robots acquiring knowledge from situated interactions involving intricate language and physical activities.

To address this challenge, this paper takes a first step towards **collaborative plan acquisition** (CPA), where humans and agents aim to communicate, learn, and infer a complete plan for joint tasks through situated dialogue. To this end, we extended `MindCraft` [Bara *et al.*, 2021], a benchmark for symmetric collaborative tasks with disparate knowledge and skills in a 3D virtual world. Specifically, we formulate a new problem for agents to predict the absent task knowledge for themselves and for their partners based on a wealth of perceptual and dialogue history. We start by annotating fine-grained dialogue moves, which capture the communicative intentions between partners during collaboration. Our hypothesis is that understanding communicative intentions plays a crucial role in ToM modeling, which, in turn, facilitate the acquisition of collaborative plans. We developed a sequence model that takes the interaction history as input and predicts the dialogue moves, the partner's mental states, and the complete plan. Our empirical results suggest that predicting the partner's missing knowledge is a more viable approach than predicting one's own. We show that explicit modeling of the partner's dialogue moves and mental states produces improved and more stable results than without.

The contributions of this work lie in that it bridges collaborative planning with situated dialogue to address how partners in a physical world can collaborate to arrive at a joint plan. In particular, it formulates a novel task on missing knowledge prediction and demonstrates that it's feasible for agents to predict their partner's missing knowledge with respect to their own partial plan. Our results have shown that, in human-AI collaboration, a more viable collaboration strategy is to infer and tell the partner what knowledge they might be missing and prompt the partner for their own missing knowledge. This strategy, if adopted by both agents, can potentially improve common ground in collaborative tasks. Our findings

---

Supplemental material along with our code can be found at https://github.com/sled-group/collab-plan-acquisition.

will provide insight for developing embodied AI agents that can collaborate and communicate with humans in the future.

2 Related Work

Our work bridges several research areas, particularly in the intersection of human-robot collaboration, planning, and theory of mind modeling.

2.1 Mixed-Initiative Planning

The essence of a collaborative task is that two participants, human or autonomous agents, pool their knowledge and skills to achieve a common goal. A mixed-initiative planning system involves a human planner and an automated planner with a goal to reduce the load and produce better plans [Queiroz Lino *et al.*, 2005]. We notice an implication that the agent's functional role is a supportive one. Work in this paradigm, called *intelligent decision support*, involves agents that range from low-level processing and/or visualization [Queiroz Lino *et al.*, 2005] to offering higher-level suggestions [Manikonda *et al.*, 2014]. Examples on this spectrum provide agents checking constraint satisfaction, eliciting user feedback on proposed plans, and ins some cases the agents are the primary decision makers [Zhang *et al.*, 2012; Gombolay *et al.*, 2015; Sengupta *et al.*, 2017]. Our desire is to have human-robot collaboration starting from an equal footing and a key goal of this is to resolve disparities in starting knowledge and abilities. We believe this can be learned from observing human-human interaction and will lead to better quality collaboration. Prior work indicates that mutual understanding and plan quality can be improved between intelligent agents through interaction [Kim and Shah, 2016], though most of the results are qualitative [Di Eugenio *et al.*, 2000], makes abstract implications [Kim and Shah, 2016], are not tractable [Grosz and Kraus, 1996]. Hybrid probabilistic generative and logic-based models that overcome incomplete information and inconsistent observations have been proposed by Kim et al. [2015]. These were successfully applied to observe natural human team planning conversations and infer the agreed-upon plan. Following these footprints, we introduce a collaborative plan acquisition task to explicitly tackle the initial disparities in knowledge and abilities of a human planner and an automated planner.

2.2 Goal and Plan Recognition

Goal recognition (GR) [Heinze, 2004; Han and Pereira, 2013] refers to the problem of inferring the goal of an agent based on the observed actions and/or their effects. On top of that, plan recognition (PR) [Kautz *et al.*, 1986; Carberry, 2001] further challenges AI agents to construct a complete plan by defining a structure with the set of observed and predicted actions that will lead to the goal [Sukthankar *et al.*, 2014; Van-Horenbeke and Peer, 2021]. We introduce a collaborative plan acquisition (CPA) task as a step forward along this line. In the CPA setting, humans and agents start both with incomplete task knowledge, communicate with each other to acquire a complete plan, and actively act in a shared environment for joint tasks. We further discuss some key benefits of our setting compared to existing work. In terms of experiment setup,

the majority of the current approaches employ plan libraries with predefined sets of possible plans [Avrahami-Zilberbrand and Kaminka, 2005; Mirsky *et al.*, 2016], or domain theories to enable plan recognition as planning (PRP) [Ramírez and Geffner, 2009; Sohrabi *et al.*, 2016], which suffer from scalability issue in complex domains [Pereira *et al.*, 2017] for high-dimensional data [Amado *et al.*, 2018]. Motivated by existing research [Rabkina *et al.*, 2020; Rabkina *et al.*, 2021], we adapt Minecraft as our planning domain, as it allows us to define agents with hierarchical plan structures and visual perception in a 3D block world that requires plan recognition from latent space. In terms of task setup, the setting of the CPA task shares the merit of active goal recognition [Shvo and McIlraith, 2020], where agents are not passive observers but are enabled to sense, reason, and act in the world. We further enable agents to communicate with their partners with situated dialogue, which is more realistic in real-world human-robot interaction. Although there exists research to integrate non-verbal communication to deal with incomplete plans in sequential plan recognition [Mirsky *et al.*, 2016; Mirsky *et al.*, 2018] and research to integrate natural language processing through parsing [Geib and Steedman, 2007], little work was done to explore language communication and dialogue processing. The CPA task introduces a more general and symmetric setting, where agents not only query their partners for missing knowledge but also actively share knowledge that their partners may be ignorant of.

2.3 Theory of Mind Modeling

As introduced by Premack and Woodruff [1978], one has a Theory of Mind (ToM) if they impute *mental states* to themselves and others. While interacting with others, Humans use their ToM to predict partners' future actions [Dennett, 1988], to plan to change others' beliefs and next actions [Ho *et al.*, 2022] and to facilitate their own decision-making [Rusch *et al.*, 2020]. In recent years, the AI community has made growing efforts to model a machine ToM to strengthen agents in human-robot interaction [Krämer *et al.*, 2012] and multiagent coordination [Albrecht and Stone, 2018]. We compare our work with representative work along this line in two dimensions. In terms of the role of the agent, prior research is largely limited to passive observer roles [Grant *et al.*, 2017; Nematzadeh *et al.*, 2018; Le *et al.*, 2019; Rabinowitz *et al.*, 2018] or as a speaker in a Speaker-Listener relationship [Zhu *et al.*, 2021]. Following a symmetric and collaborative setup [Bara *et al.*, 2021; Sclar *et al.*, 2022], we study ToM modeling in agents that actively interact with the environment and engage in free-form situated communication with a human partner. In terms of task formulation, machine ToM has been typically formulated as inferring other agents' beliefs [Grant *et al.*, 2017; Nematzadeh *et al.*, 2018; Le *et al.*, 2019; Bara *et al.*, 2021], predicting future actions [Rabinowitz *et al.*, 2018], generating pragmatic instructions [Zhu *et al.*, 2021], and gathering information [Sclar *et al.*, 2022]. None of these formulations were able to explicitly assess how well can AI agents use their machine ToM to complete partial plans through situation communication with their collaborators, as humans usually do in real-world interactions. To this end, we extended the prob-

lem formulation in [Bara *et al.*, 2021] to a collaborative plan acquisition task, where humans and agents try to learn and communicate with each other to acquire a complete plan for joint tasks.

3 Background: ToM for Collaborative Tasks

Our work is built upon `MindCraft`, a benchmark and platform developed by [Bara *et al.*, 2021] for studying ToM modeling in collaborative tasks. We first give a brief introduction to this benchmark and then illustrate how our work differs from `MindCraft`.[1]

The `MindCraft` platform supports agents to complete collaborative tasks through situated dialogue in a 3D block world, with rich mental state annotations. As shown in Figure 1, two agents are co-situated in a shared environment and their joint goal is to create a specific material. There are two macro-actions: (1) creating a block and (2) combining two existing blocks to create a new block. These macro-actions are made up of atomic actions that agents may perform in-game, *e.g.*, navigation, jumping, and moving blocks around. During gameplay, the two agents are each given a partial plan in the form of a directed AND-graph. These partial plans are incomplete in the sense that the agents cannot complete the task by themselves individually by following the plan. The two players will need to communicate with each other through the in-game chat so that they can coordinate with each other to complete the joint task. An example dialogue session between the players to communicate the plan is shown in Figure 1.

As an initial attempt, Bara et al. [2021] formulated ToM modeling as three tasks that predict a partner's mental state:

- **Task intention**: predict the sub-goal that the partner is currently working on;
- **Task status**: predict whether the partner believes a certain sub-goal is completed and by whom;
- **Task knowledge**: predict whether the partner knows how to achieve a sub-goal, *i.e.*, all the incoming edges of a node;

A baseline was almost implemented that takes perceptual observation and interaction history for these prediction tasks and reported results in [Bara *et al.*, 2021]. For the remainder of this paper, we use ToM tasks to refer to these three tasks introduced in the original paper.

It's important to note that, although we use the `MindCraft` benchmark, our work here has several significant differences. First and foremost, while `MindCraft` studies ToM, it primarily focused on inferring other agents' mental states, and has not touched upon collaborative plan acquisition. How humans and agents communicate, learn, and infer a complete plan for joint tasks through situated dialogue is the new topic we attempt to address in this paper. Second, `MindCraft` mostly focuses on ToM modeling and only provides ground-truth labels for the three tasks described above. As communicative intentions play an important role in coordinating activities between partners, we added additional

annotations for dialogue moves (as shown in Figure 1). We investigate if the incorporation of dialogue moves would benefit mental state prediction and plan acquisition.

4 Collaborative Plan Acquisition

In a human-AI team like that in `MindCraft`, humans and AI agents may have insufficient domain knowledge to derive a complete plan, thus suffering from an incomplete action space to execute a complete plan. It's therefore important for an agent to predict what knowledge is missing for themselves and for their partners, and proactively seek/share that information so the team can reach a common and complete plan for the joint goal. We start by formalizing the plan representation, followed by a description of the collaborative plan acquisition problem.

4.1 Task Formulation

Definition 1 (Joint and Partial Plan). *We represent a joint plan $\mathcal{P} = (V, E)$ as a directed AND-graph, where the nodes V denote sub-goals and the edges E denote temporal constraints between the subgoals. As a directed AND-graph, all of the children sub-goals of an AND-node must be satisfied in order to perform the parent. A partial plan $\tilde{\mathcal{P}} = (V, \tilde{E})$ is a subgraph of $\mathcal{P}$ with a shared set of nodes V but only share a subset of edges $\tilde{E} \subseteq E$.*

An example of a complete plan graph in `MindCraft` is shown in Figure 2. Each plan contains a joint goal, and the rest of the nodes denotes fact landmarks [Hoffmann *et al.*, 2004], *i.e.*, sub-goals that must be achieved at some point along all valid execution.

We consider a pair of collaborative agents with a joint plan $\mathcal{P}$. To account for the limited domain knowledge, an agent i has an initial partial plan $\mathcal{P}_i = (V, E_i), E_i \subseteq E$, which is a subgraph of $\mathcal{P}$. As shown in Figure 2, the complete plan and partial plans share the same set of nodes (*i.e.*, the same V). The agent only has access to its own knowledge, which might be shared (denoted as blue arrows) or not (denoted as green arrows). Its missing knowledge (denoted as grey arrows) is a set of edges $\overline{E}_i = E \backslash E_i$. We assume, in this work, the collaborative planning problem is solvable for the collaborative agents, *i.e.*, $\bigcup E_i = E$. In order for the agents to solve the problem, agents can communicate with their partners by sending a natural language message, which in turn helps them to acquire a complete plan graph. We define this collaborative plan acquisition problem formally as follows.

Definition 2 (Collaborative Plan Acquisition Problem). *In a collaborative plan acquisition problem with a joint plan $\mathcal{P}$, an agent i and its collaborative partner j start with partial plans $\mathcal{P}_i = (V, E_i)$ and $\mathcal{P}_j = (V, E_j)$. At each timestamp t, the agent i has access to a sequence of up-to-date visual observations $O_i^{(t)}$ and dialogue history $D^{(t)}$. The problem is for agent i's to acquire its own missing knowledge $\overline{E}_i = E \backslash E_i$ and the partner j's missing knowledge $\overline{E}_j = E \backslash E_j$.*

To solve a collaborative plan acquisition problem, agent i needs to address two tasks.

---

[1]The original dataset consists of 100 dialogue sessions. We used the platform and collected 60 additional sessions to increase the data size for our investigation under the IRB (HUM00166817) approved by the University of Michigan.
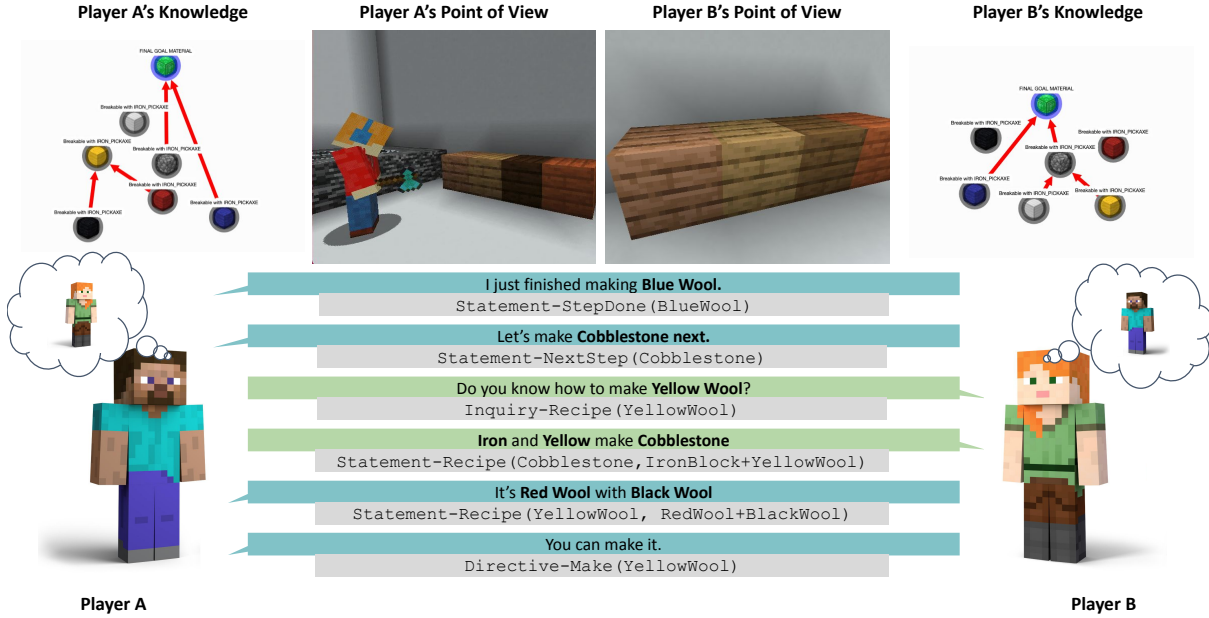
Figure 1: An example dialogue history between two partners to complete a joint goal in `MindCraft`. Each player is given a partial plan. They communicate with each other to form a complete plan for the goal. Each utterance is annotated with a dialogue move that describes the communicative intention.
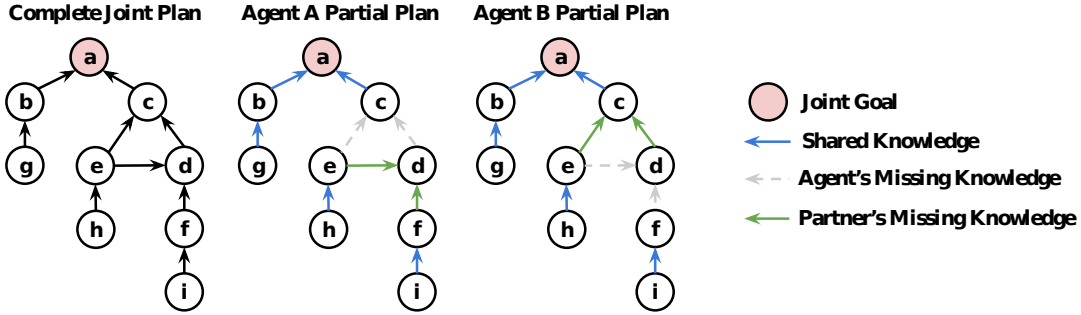

Figure 2: An example of the plan graphs. From left to right, we illustrate a complete joint plan, a partial plan for Agent A, and another partial plan for Agent B. The plan graphs all contain the same set of nodes, with a *joint goal a* on top and other nodes representing *fact landmarks*. In a collaborative plan acquisition problem, an agent is tasked to infer its own missing knowledge and its partner's missing knowledge.

Task 1: Inferring own missing knowledge. For Task 1, a solution is a set of missing knowledge $\overline{E}_i = E \backslash E_i$. Agent i needs to infer its own missing knowledge by identifying the missing edges in its partial plan from the complete joint plan, *i.e.*, $P(e \in E \backslash E_i \mid O_i^{(t)}, D_i^{(t)}), \forall e \in V^2 \backslash E_i$ at time t. Note that V^2 refers to a complete graph, instead of a complete joint plan. For example, as shown in Figure 2, among all the missing edges in Agent A's partial plan, we hope Agent A would correctly predict that the edge $d \rightarrow c$ and the edge $e \rightarrow c$ are missing in their own plan. Recovering those edges leads to a complete joint plan.

Task 2: Inferring partner's missing knowledge. For Task 2, a solution is a set of missing knowledge $\overline{E}_j = E \backslash E_j$. Agent i predicts what edges in i's partial plan which might be missing in its partner j's partial plan, *i.e.*, $P(e \in E \backslash E_j \mid O_i^{(t)}, D_i^{(t)}), \forall e \in E_i$ at time t. In the example in

Figure 2, Agent A should select the edges $e \rightarrow d$ and $f \rightarrow d$ from their own partial plan as being absent from their partner's plan. If the agent can correctly predict which edges are missing for their partner, the agent can proactively communicate to their partner and therefore help their partner acquire a complete task plan. If agents can predict what each other is missing and proactively share that knowledge, then both agents will be able to reach a common understanding of the complete joint plan.

We note that the **Task Knowledge** in ToM tasks is different from the **Task 2** we propose. In Task Knowledge, the model is probed whether *one* piece of knowledge that might be unknown itself is known by the partner. In Task 2, the model needs to predict, for *each* piece of the agent's own knowledge, whether the partner shares it or not.
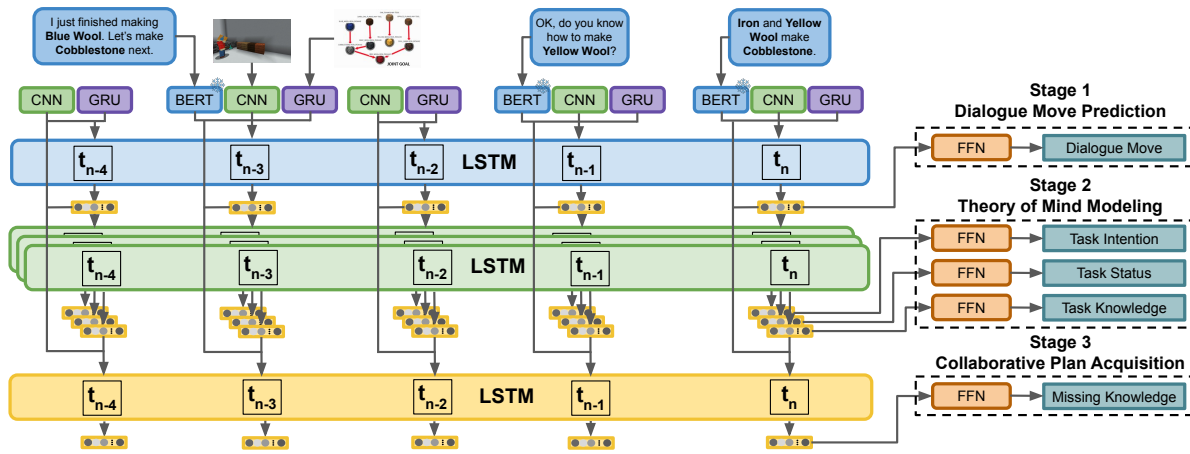
Figure 3: The theory of mind (ToM) model consists of a base sequence model taking in as input representations for dialogue (D) when available), visual observation of the environment (O), and the partial plan available to the agent. The model can be configured to take optional inputs as latent representations from the frozen mental state prediction models and the dialogue move representation for dialogue exchanges.

4.2 Dialogue Moves in Coordination of Plans

While the ToM tasks have captured the partner's *task intention*, another important dimension of intention, the *communicative intention*, was neglected in the original `MindCraft` paper. This can be captured by *dialogue moves*, which are sub-categories of dialogue acts that guide the conversation and update the information shared between the speakers [Traum and Larsson, 2003; Marge *et al.*, 2020]. To this end, we introduce a dialogue move prediction task to better understand the dialogue exchanges between the agent and its partner in a collaborative planning problem. We build our move schema from the set of dialogue acts described in [Stolcke *et al.*, 2000], out of which we keep a relevant subset. We expand the `Directive`, `Statement`, and `Inquiry` categories for domain-specific uses in the `MindCraft` collaborative task. For these three categories, we introduce parameters that complete the semantics of dialogue moves. These parameters serve the purpose of grounding the dialogue to the partial plan graph given to the player. We show all dialogue moves with parameters in the supplemental material.

4.3 Computational Model

End-to-end baseline. We start by introducing a straightforward end-to-end approach to address both Task 1 and Task 2, similar to the model in [Bara *et al.*, 2021] for consistency. More specifically, the baseline model processes the dialogue input with a frozen language model [Devlin *et al.*, 2019], the video frames with a convolutional neural network, and the partial plan with a gated recurrent unit [Chung *et al.*, 2014]. The sequences of visual, dialogue, and plan representations are fed into an LSTM [Hochreiter and Schmidhuber, 1997]. The predictions of the missing knowledge of the agent and its partner are decoded with feed-forward networks.

Augmented models. One important research question we seek to address is whether explicit treatment of the partner's mental states and dialogue moves would benefit collaborative plan acquisition. As shown in Figure 3, we develop a multi-stage augmented model that first learns to predict the dialogue moves, then attempts to model a theory of mind, and finally learns to predict the missing knowledge of itself and of its partner respectively. The model processes each input modality and time series with the same architectural design as the baseline. At Stage 1, the models predict the dialogue moves of the partner. At Stage 2, we freeze the model pre-trained in Stage 1 and concatenate their output latent representations to augment the input sequences of visual, dialogue, and plan representations. The models predict the mental states of the partner, with each LSTM sequence model dedicated to task intention, task status, and task knowledge. At Stage 3, we freeze the models pre-trained in Stages 1 and 2 and task the model to predict the missing knowledge with similar LSTM and feed-forward networks.

5 Empirical Studies

In this section, we first examine the role of dialogue moves in three ToM tasks and further discuss how dialogue moves and ToM modeling influence the quality of collaborative plan acquisition.

5.1 Role of Dialogue Moves in ToM Tasks

The performance of the mental state prediction models, as presented in `MindCraft`, shows low performance in the multimodal setting. We begin by confirming the effectiveness of dialogue moves, by evaluating if they help to improve these ToM tasks proposed in [Bara *et al.*, 2021]. The same setting as described by the Stage 1 and 2. We show results in Table 1, which compares the performance of the baseline model with the model augmented with dialogue moves. We observe a significant increase in performance when using dialogue moves. Furthermore, for the task of predicting the partner's task knowledge, we observe that the augmented model approaches the average human performance.[2]

---

[2]The average human performance measure was provided in the original `MindCraft` [Bara *et al.*, 2021].

The best-performing models for every ToM task are used to produce the latent representations for our subsequent tasks.

Task Status			
Modalities	**w/o Dlg Moves**	**w/ Dlg Moves**	**Human**
None	N/A	**56.0**±0.8	67.0
D	45.8±3.0	54.6±1.1	67.0
D+O	32.7±1.2	**59.3**±1.0	67.0
O	53.7±1.1	**59.3**±1.7	67.0
Task Knowledge			
Modalities	**w/o Dlg Moves**	**w/ Dlg Moves**	**Human**
None	N/A	**54.7**±2.5	58.0
D	45.3±1.3	**56.2**±1.9	58.0
D+O	48.3±1.1	**57.6**±1.0	58.0
O	49.4±2.5	**56.4**±2.5	58.0
Task Intention			
Modalities	**w/o Dlg Moves**	**w/ Dlg Moves**	**Human**
None	N/A	**14.9**±1.5	46.0
D	3.0±0.6	**12.1**±1.0	46.0
D+O	6.2±0.6	**13.5**±0.6	46.0
O	6.6±1.1	**13.8**±1.7	46.0

Table 1: Performance on the three ToM tasks with or without using dialogue moves. `None` means only dialogue moves are used for prediction. `D` stands for text in dialogue; `O` stands for visual observation of the activities in the environment; `D+O` both text and visual observation. Highlighted values are statistically significant with $P < 0.01$ compared to the best model without dialogue moves for the given task.

5.2 Results for Collaborative Plan Acquisition

We now present the empirical results for collaborative plan acquisition, *i.e.*, inferring one's own missing knowledge and inferring the partner's missing knowledge at the end of each session. We use the following metrics to evaluate the performance on these tasks:

- **Per edge F1 score**, computed by aggregating all edges across tasks. It is meant to evaluate the model's ability to predict whether an edge is missing in a partial plan.

- **Per task F1 score**, computed as the average of F1 scores within a dialogue session. It is meant to evaluate the model's average performance across sessions.

Task 1: Inferring own missing knowledge. The performance is shown in Table 2a. Overall, we found the models underperform across all configurations, meaning that inferring one's own missing knowledge turns out to be a difficult task. We believe this is due to the sparsity of the task graph. Since the space of possible edges to be predicted is large (as the agent needs to consider every possible link between two nodes), the link prediction becomes notoriously challenging. Better solutions will be needed for this task in the future.

Task 2: Inferring partner's missing knowledge. Table 2b shows the performance which is across the board more than 70% F1. This means that this task is more approachable compared to inferring one's own missing knowledge. Table 2b also compares various combinations of augmented models, with different augmentations available from Stage 1 (*e.g.*, latent representations of dialogue moves and other mental states). While all combinations lead to increased performance, we found in general that incorporating dialogue

a) Performance on inferring agent's own knowledge.

Task Status	Task Know.	Task. Int.	Dlg. Move	Per Edge F1 Score	Per Task F1 Score
				17.0 ± 0.2	19.8 ± 1.0
	X			19.0 ± 2.5	21.1 ± 1.6
		X		**21.0** ± 0.7	22.2 ± 2.2
	X	X		**19.6** ± 1.4	21.4 ± 1.7
	X		X	**20.1** ± 1.4	22.1 ± 1.2
		X	X	**19.8** ± 1.7	21.7 ± 1.8
X	X	X	X	17.4 ± 0.1	20.0 ± 1.9

b) Performance on inferring the player's partner's plan.

Task Status	Task Know.	Task. Int.	Dlg. Move	Per Edge F1 Score	Per Task F1 Score
				71.3 ± 1.1	68.8 ± 3.1
	X			**74.4** ± 0.3	73.6 ± 1.4
		X		**74.5** ± 1.4	73.8 ± 3.0
	X	X		**74.4** ± 1.4	73.5 ± 2.3
	X		X	**74.3** ± 0.7	**73.5** ± 1.3
		X	X	**75.0** ± 1.0	**74.7** ± 2.2
X	X	X	X	**73.5** ± 0.5	72.1 ± 1.8

Table 2: Highlighted values are statistically significant with $P < 0.01$ compared to the base model without augmentation.

moves and task intention has the highest performance. This finding confirms the importance of intention prediction in plan coordination, from both the communication level and task level.

5.3 Cross-time Analysis

As Task 2 is more approachable with more reasonable performance, we look further into the performance of the models as the interaction progresses. In this section, we look at not only the latent representation at the end of an interaction but also throughout the interaction, to predict the partner's missing knowledge. We find that as the interaction progresses, in absolute time, there is an upward trend concerning both per edge F1 score (Figure 4a), as well as concerning per task F1 score (Figure 4b). This is not surprising. As models have access to richer interaction history, which enables them to make more reliable predictions of the partner's missing knowledge. Furthermore, we observe that the augmented models improve the performance across time, confirming the results we obtained in the previous section.

We are also interested in the stability of system prediction over time, which is an important feature if these agents are ever employed in real interaction with humans. We introduce two metrics for this purpose:

- **Number of mind changes**, which is the number of times that the model changes its prediction of whether an edge is missing from its partner's knowledge.

- **Accumulated absolute confidence changes**, which adds up the absolute value of the changes in the prediction probability between timestamps.

Figure 5a shows the average number of mind changes and Figure 5b shows the accumulated absolute confidence changes as the interaction progresses. We observe that the augmented models show significantly fewer changes in prediction and a lower change in prediction confidence, as compared with the base model. These results have shown that the base model is more inclined to "change its mind" in prediction as interaction proceeds, the augmented models that take
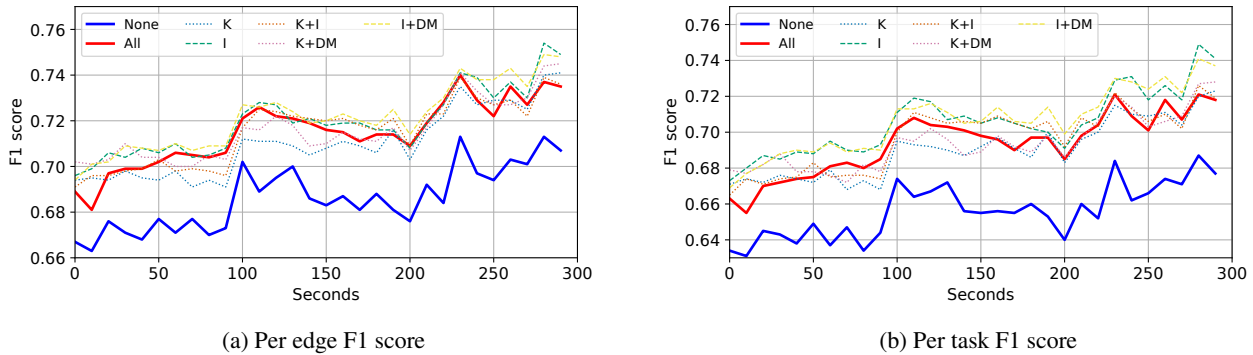
(a) Per edge F1 score

(b) Per task F1 score

Figure 4: Model performance as the interaction progresses in absolute time. The abbreviations are S - Task Status, K - Task Knowledge, I - Task Intent, and DM - Dialogue Moves.
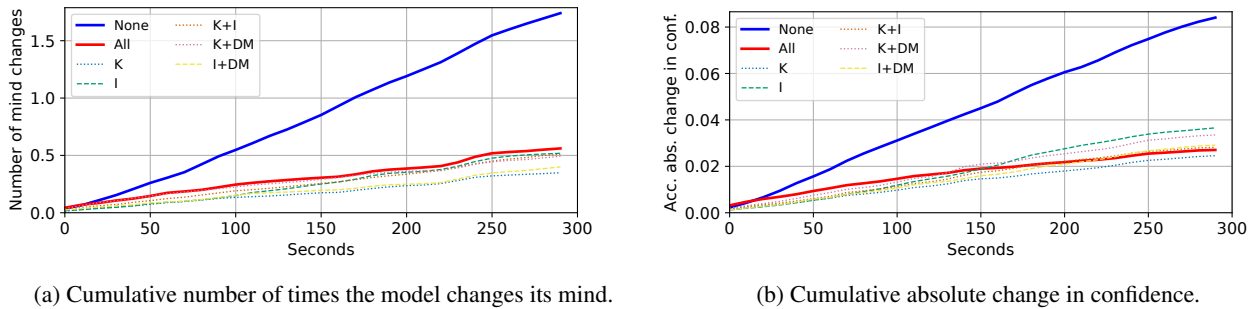

(a) Cumulative number of times the model changes its mind.

(b) Cumulative absolute change in confidence.

Figure 5: The average change in model prediction over each edge as the interaction progresses in absolute time. The cause for the curve's dips is that not all interactions are of equal length. The further in time, the fewer interactions there are that have data at that time step. The abbreviations are S - Task Status, K - Task Knowledge, I - Task Intent, and DM - Dialogue Moves.

into account of partner's mental states and dialogue moves are more stable in their prediction throughout the interaction.

6   Discussion and Conclusion

In this work, we address the challenge of collaborative plan acquisition in human-agent collaboration. We extend the `MindCraft` benchmark and formulate a problem for agents to predict missing task knowledge based on perceptual and dialogue history, focusing on understanding communicative intentions in Theory of Mind (ToM) modeling. Our empirical results highlight the importance of predicting the partner's missing knowledge and explicitly modeling their dialogue moves and mental states. A promising strategy for effective collaboration involves inferring and communicating missing knowledge to the partner while prompting them for their own missing knowledge. This collaborative approach holds the potential to improve decision-making when both agents actively engage in its implementation. The findings have implications for the development of embodied AI agents capable of seamless collaboration and communication with humans. Specifically, by predicting their partner's missing knowledge and actively sharing that information, these agents can facilitate a shared understanding and successful execution of joint tasks. The future efforts following this research could explore and refine this collaborative strategy.

This work presents our initial results. It also has several

limitations. The current setup assumes shared goals and a single optimal complete plan without alternatives, neglecting the complexity that arises from the absence of shared goals and the existence of alternative plans. Our motivation for controlling the form of partial plans and the predetermined complete plan is to enable a systematic focus on modeling and evaluating plan coordination behaviors. Although our current work is built on the `MindCraft` dataset where partial plans are represented by AND-graphs, the problem formulation can be potentially generalized to multiple AND-OR-graphs. Future research could explore approaches that incorporate AND-OR-graphs to account for alternative paths to achieving joint goals.

Additionally, our present study focuses on a dyadic scenario, employing human-human collaboration data to study collaborative plan acquisition. Since AI agents typically have limited visual perception and reasoning abilities compared to their human counterparts, the communication discourse is expected to exhibit increased instances of confirmations, repetitions, and corrections. How to effectively extend the models trained on human-human data to human-agent collaboration remains an important question. With the emergence of large foundation models [Bommasani et al., 2021], our future work will incorporate these models into our framework to facilitate situated dialogue for collaborative plan acquisition. We will further conduct experiments and evaluate the efficacy of these models in more complex human-agent collaborations.

Acknowledgments

References

[Albrecht and Stone, 2018] Stefano V Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95, 2018.

[Amado *et al.*, 2018] Leonardo Rosa Amado, João Paulo Aires, RAMON FRAGA PEREIRA, Maurício Cecílio Magnaguagno, Roger Leitzke Granada, and Felipe Rech Meneguzzi. Lstm-based goal recognition in latent space. In *ICML/IJCAI/AAMAS 2018 Workshop on Planning and Learning (PAL-18), 2018, Suécia.*, 2018.

[Avrahami-Zilberbrand and Kaminka, 2005] Dorit Avrahami-Zilberbrand and Gal A Kaminka. Fast and complete symbolic plan recognition. In *IJCAI*, pages 653–658, 2005.

[Bara *et al.*, 2021] Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[Bommasani *et al.*, 2021] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[Carberry, 2001] Sandra Carberry. Techniques for plan recognition. *User modeling and user-adapted interaction*, 11(1):31–48, 2001.

[Chung *et al.*, 2014] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[Dennett, 1988] Daniel C Dennett. Précis of the intentional stance. *Behavioral and brain sciences*, 11(3):495–505, 1988.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[Di Eugenio *et al.*, 2000] Barbara Di Eugenio, Pamela W Jordan, Richmond H Thomason, and JOHANNA D MOORE. The agreement process: An empirical investigation of human–human computer-mediated collaborative dialogs. *International Journal of Human-Computer Studies*, 53(6):1017–1076, 2000.

[Geib and Steedman, 2007] Christopher W Geib and Mark Steedman. On natural language processing and plan recognition. In *IJCAI*, volume 2007, pages 1612–1617, 2007.

[Gombolay *et al.*, 2015] Matthew C Gombolay, Reymundo A Gutierrez, Shanelle G Clarke, Giancarlo F Sturla, and Julie A Shah. Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. *Autonomous Robots*, 39(3):293–312, 2015.

[Grant *et al.*, 2017] Erin Grant, Aida Nematzadeh, and Thomas L Griffiths. How can memory-augmented neural networks pass a false-belief task? In *CogSci*, 2017.

[Grosz and Kraus, 1996] Barbara J Grosz and Sarit Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357, 1996.

[Han and Pereira, 2013] The Anh Han and Luís Pereira. State-of-the-art of intention recognition and its use in decision making: A research summary. *AI Communications*, 26(2):237–246, 2013.

[Heinze, 2004] Clint Heinze. Modelling intention recognition for intelligent agent systems. Technical report, Defence Science and Technology Organisation Salisbury (Australia) Systems . . . , 2004.

[Ho *et al.*, 2022] Mark K Ho, Rebecca Saxe, and Fiery Cushman. Planning with theory of mind. *Trends in Cognitive Sciences*, 2022.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Hoffmann *et al.*, 2004] Jörg Hoffmann, Julie Porteous, and Laura Sebastia. Ordered landmarks in planning. *Journal of Artificial Intelligence Research*, 22:215–278, 2004.

[Kautz *et al.*, 1986] Henry A Kautz, James F Allen, et al. Generalized plan recognition. In *AAAI*, volume 86, page 5. Philadelphia, PA, 1986.

[Kim and Shah, 2016] Joseph Kim and Julie A Shah. Improving team's consistency of understanding in meetings. *IEEE Transactions on Human-Machine Systems*, 46(5):625–637, 2016.

[Kim *et al.*, 2015] Been Kim, Caleb M Chacha, and Julie A Shah. Inferring team task plans from human meetings: A generative modeling approach with logic-based prior. *Journal of Artificial Intelligence Research*, 52:361–398, 2015.

[Krämer *et al.*, 2012] Nicole C Krämer, Astrid von der Pütten, and Sabrina Eimler. Human-agent and human-robot interaction theory: Similarities to and differences from human-human interaction. In *Human-computer interaction: The agency perspective*, pages 215–240. Springer, 2012.

[Le *et al.*, 2019] Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. In *EMNLP-IJCNLP 2019*, pages 5872–5877, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Manikonda *et al.*, 2014] Lydia Manikonda, Tathagata Chakraborti, Sushovan De, Kartik Talamadupula, and

Subbarao Kambhampati. Ai-mix: using automated planning to steer human workers towards better crowd-sourced plans. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

[Marge *et al.*, 2020] Matthew Marge, Felix Gervits, Gordon Briggs, Matthias Scheutz, and Antonio Roque. Let's do that first! a comparative analysis of instruction-giving in human-human and human-robot situated dialogue. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*, 2020.

[Mirsky *et al.*, 2016] Reuth Mirsky, Roni Stern, Ya'akov Gal, and Meir Kalech. Sequential plan recognition. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 401–407, 2016.

[Mirsky *et al.*, 2018] Reuth Mirsky, Roni Stern, Kobi Gal, and Meir Kalech. Sequential plan recognition: An iterative approach to disambiguating between hypotheses. *Artificial Intelligence*, 260:51–73, 2018.

[Nematzadeh *et al.*, 2018] Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[Pereira *et al.*, 2017] Ramon Pereira, Nir Oren, and Felipe Meneguzzi. Landmark-based heuristics for goal recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[Premack and Woodruff, 1978] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978.

[Queiroz Lino *et al.*, 2005] Natasha Queiroz Lino, Austin Tate, and Yun-Heh Chen-Burger. Semantic support for visualisation in collaborative ai planning. In Susanne Biundo, Karen Myers, and Kanna Rajan, editors, *Proceedings of the Workshop on The Role of Ontologies in Planning and Scheduling*. AAAI Press, June 2005. The Fifteenth International Conference on Automated Planning and Scheduling, ICAPS.

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

[Rabkina *et al.*, 2020] Irina Rabkina, Pavan Kathnaraju, Mark Roberts, Jason Wilson, Kenneth Forbus, and Laura Hiatt. Recognizing the goals of uninspectable agents. In *Proceedings of the Conference on Advances in Cognitive Systems (Palo Alto, CA:)*, pages 1–8, 2020.

[Rabkina *et al.*, 2021] Irina Rabkina, Pavan Kantharaju, Jason R Wilson, Mark Roberts, and Laura M Hiatt. Evaluation of goal recognition systems on unreliable data and uninspectable agents. *Frontiers in Artificial Intelligence*, 4, 2021.

[Ramírez and Geffner, 2009] Miquel Ramírez and Hector Geffner. Plan recognition as planning. In *Twenty-First international joint conference on artificial intelligence*, 2009.

[Rusch *et al.*, 2020] Tessa Rusch, Saurabh Steixner-Kumar, Prashant Doshi, Michael Spezio, and Jan Gläscher. Theory of mind and decision science: towards a typology of tasks and computational models. *Neuropsychologia*, 146:107488, 2020.

[Sclar *et al.*, 2022] Melanie Sclar, Graham Neubig, and Yonatan Bisk. Symmetric machine theory of mind. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 19450–19466, 17–23 Jul 2022.

[Sengupta *et al.*, 2017] Sailik Sengupta, Tathagata Chakraborti, Sarath Sreedharan, Satya Gautam Vadlamudi, and Subbarao Kambhampati. Radar-a proactive decision support system for human-in-the-loop planning. In *2017 AAAI Fall Symposium Series*, 2017.

[Shvo and McIlraith, 2020] Maayan Shvo and Sheila A McIlraith. Active goal recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9957–9966, 2020.

[Sohrabi *et al.*, 2016] Shirin Sohrabi, Anton V Riabov, and Octavian Udrea. Plan recognition as planning revisited. In *IJCAI*, pages 3258–3264. New York, NY, 2016.

[Stolcke *et al.*, 2000] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373, 2000.

[Sukthankar *et al.*, 2014] Gita Sukthankar, Christopher Geib, Hung Bui, David Pynadath, and Robert P Goldman. *Plan, activity, and intent recognition: Theory and practice*. Newnes, 2014.

[Traum and Larsson, 2003] David R Traum and Staffan Larsson. The information state approach to dialogue management. In *Current and new directions in discourse and dialogue*, pages 325–353. Springer, 2003.

[Van-Horenbeke and Peer, 2021] Franz A Van-Horenbeke and Angelika Peer. Activity, plan, and goal recognition: A review. *Frontiers in Robotics and AI*, 8:643010, 2021.

[Zhang *et al.*, 2012] Haoqi Zhang, Edith Law, Rob Miller, Krzysztof Gajos, David Parkes, and Eric Horvitz. Human computation tasks with global constraints. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 217–226, 2012.

[Zhu *et al.*, 2021] Hao Zhu, Graham Neubig, and Yonatan Bisk. Few-shot language coordination by modeling theory of mind. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12901–12911, 18–24 Jul 2021.