Infer Human's Intentions Before Following Natural Language Instructions

Yanming Wan[1], Yue Wu[1], Yiping Wang[1], Jiayuan Mao[2*], Natasha Jaques[1*]

[1]University of Washington, Seattle, WA 98195
[2]MIT CSAIL, Cambridge, MA 02139
{ymwan, nj}@cs.washington.edu, jiayuanm@mit.edu

Abstract

For AI agents to be helpful to humans, they should be able to follow natural language instructions to complete everyday cooperative tasks in human environments. However, real human instructions inherently possess ambiguity, because the human speakers assume sufficient prior knowledge about their hidden goals and intentions. Standard language grounding and planning methods fail to address such ambiguities because they do not model human internal goals as additional partially observable factors in the environment. We propose a new framework, Follow Instructions with Social and Embodied Reasoning (FISER), aiming for better natural language instruction following in collaborative embodied tasks. Our framework makes explicit inferences about human goals and intentions as intermediate reasoning steps. We implement a set of Transformer-based models and evaluate them over a challenging benchmark, HandMeThat. We empirically demonstrate that using social reasoning to explicitly infer human intentions before making action plans surpasses purely end-to-end approaches. We also compare our implementation with strong baselines, including Chain of Thought prompting on the largest available pre-trained language models, and find that FISER provides better performance on the embodied social reasoning tasks under investigation, reaching the state-of-the-art on HandMeThat.

Code — https://github.com/Simon-Wan/FISER
Extended version — https://arxiv.org/abs/2409.18073

Introduction

Building AI assistants that can interact with people in a shared environment and follow their instructions would unlock assistive robotics and free up domestic labor. Toward this broad goal, we need to address the problem of "translating" realistic natural language instructions into actions executable by robots. The conventional way that people formulate this problem is grounded language learning, which aims at mapping abstract natural language phrases to concretely executable actions. However, these approaches miss an important component of many human-robot collaborative tasks, which is that the language humans tend to use in everyday scenarios is inherently ambiguous. Human speakers assume that listeners possess prior knowledge, leading them to omit certain information for efficiency (Grice 1975; Sperber and Wilson 1986; Clark 1996; Dennett 1987; Gergely et al. 1995). Resolving this ambiguity depends on leveraging other sources of information (e.g., human internal goals and historical actions) that are partially observable to the robot.

Consider the example shown in Fig. 1, where a human is tidying up a room. In the middle of her actions, she asks a robot for help, saying *"Could you pass that from the sofa?"* This instruction does not appear to be solvable without further information about the person's underlying intention. While such internal mental states are not directly observed, agents can infer them from human's past actions. Specifically, if the robot can observe that in previous steps, the human put several books into a box one by one, it can infer that she intends to use that box to store all the books. Based on this guess, the robot can check if there are any remaining books on the sofa and then hand them to the person.

Generally speaking, the ambiguity in the instruction mainly arises from two aspects. First, the human assumes sufficient prior knowledge about her hidden intentions (Dennett 1987; Gergely et al. 1995), which is based on the common sense knowledge that people tend to group similar items together when tidying up, and the observation that the human is gathering books. Second, people make trade-offs between accuracy and efficiency of communication (Grice 1975; Sperber and Wilson 1986; Clark 1996). This leads to the challenge of building AI agents that can follow efficient, ambiguous speech that people naturally adopt when giving directions.

We consider the case where there's a human and a robot collaborating in a shared environment. The human is working on some tasks, and specifies a sub-task for the robot to help with by giving a natural language instruction. Past methods (e.g., language grounding) attempt to directly complete the specified command from the given instructions, since the human only acts as a disembodied issuer of instructions and is not another active agent in their environments. The human, as another partially observable factor in the environment, has been overlooked. In this paper, we present a new framework, Follow Instructions with Social and Embodied Reasoning (FISER), which suggests that we should introduce the human's intention as explicit variables for the model to draw inferences about. By leveraging this structure, our framework opts to decompose the problem into two parts – **social reason-**

*These authors contributed equally.

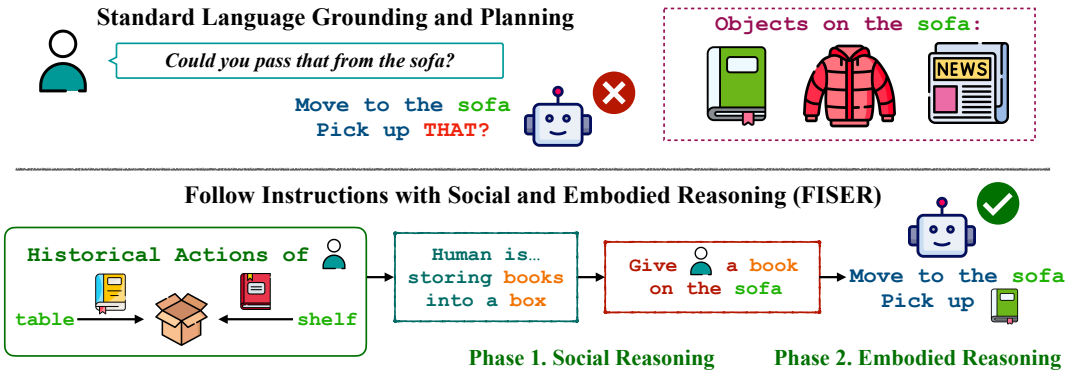

Figure 1: An example scenario where the human's natural language instruction ("*Could you pass that from the sofa?*") is inherently ambiguous. Standard language grounding and planning methods fail to resolve ambiguity. We propose FISER, which explicitly reasons about human's internal intentions as intermediate steps. The robot disamiguates the instruction into a concrete robot-understandable task in the social reasoning phase (Phase 1), and then accomplishes the grounded planning in the embodied reasoning phase (Phase 2). We further propose an optional enhancement to Phase 1 by explicitly recognize the human's overall plan first, and then infer what the human want the robot to do.

ing and **embodied reasoning**. Specifically, **social reasoning** is aimed at predicting the sub-task for which the human is asking for assistance, which can be inferred from the context of both the instruction and the observed historical actions of the person in the shared environment. After grounding the instructions into robot-understandable tasks, the robot can then do planning and interact with the environment, in a separate **embodied reasoning** phase. To further enhance the model's ability to follow ambiguous instructions, we propose to explicitly add an extra plan recognition stage, where a set of logical predicates is used to help with inferring the human's overall plan. We implement a Transformer-based model trained in a supervised learning manner to predict specified sub-tasks (and the human's underlying plan) at intermediate layers. This step-by-step approach distinctly differs from the more commonly employed end-to-end methods in previous works.

Overall, the key insight of our method is that separating social and embodied reasoning by explicitly modeling the human's intentions can significantly improve performance when following ambiguous natural language instructions. To test this hypothesis, we evaluate our models on a challenging benchmark, HandMeThat (HMT) (Wan, Mao, and Tenenbaum 2022), which involves ambiguous instruction following tasks in a text-based household environment. HMT contains a large number of physical objects and valid actions in each episode, as well as an enormous human goal space. We find that these properties make HMT challenging even for the largest state-of-the-art large language models (LLMs). As a competitive baseline, we also design a Chain-of-Thought (CoT) approach to prompt GPT-4 based on our framework. There are two findings from the experimental results. First, models which separate social and embodied reasoning using the FISER framework outperform end-to-end reasoning, in both Transformer-based models and CoT-prompted LLMs, which indicates that explicitly doing intermediate reasoning about human intentions is beneficial. Second, training small-scale models from scratch on this task outperforms our most sophisticated CoT prompting methods for large pre-trained LLMs, indicating that pretraining and domain-specific prompts are insufficient for LLMs to perform well on the challenging social and embodied reasoning tasks.

To summarize, the contributions of this paper are to propose the FISER framework, which performs instruction following by first using social reasoning and additional context to disambiguate what the human is asking, before using embodied reasoning to decide what actions to take to complete the task. We further introduce a human plan recognition stage to enhance social reasoning abilities when tasks are particularly complex or ambiguous. We empirically demonstrate that our FISER models show 64.5% success rate on the test set on average, achieving the state-of-the-art on HMT benchmark.

Related Work

Grounded language learning. In order for AI to be useful to people in our homes and natural environments, non-experts need to be able to communicate with AI agents using natural language. This issue has long captured the attention of researchers (Winograd 1972; Siskind 1994), and the primary challenge involves mapping natural language to concrete meanings within the physical environment. Several studies explore language-conditioned task completion in specific environments (Shridhar et al. 2020; Suglia et al. 2021; Kojima, Suhr, and Artzi 2021). With the emergence of LLMs, many works discussed grounding language by leveraging pre-trained LLMs (Blukis et al. 2021; Nair et al. 2022; Zellers et al. 2021; Zhi-Xuan et al. 2024; Min et al. 2024). A prominent example is SayCan (Ahn et al. 2022), which proposed extracting the knowledge in LLMs by using them to score the likelihood that a subtask available to the robot will help complete a high-level instruction. Although the above studies may incorporate common sense reasoning about language as

well as information within the physical environment, their instructions explicitly express human intentions. For example, the most ambiguous instruction solved by SayCan is, "*I spilled my coke, can you bring me something to clean it up?*" where the ambiguity can still be easily resolved given that *the sponge* is the only cleaning tool in the environment. In contrast, we address the problem that realistic human instructions omit certain information for efficiency, making them much more ambiguous, and necessitating inferring human intentions to fill in the gaps.

Collaborative communication. We consider the case where the human and the robot are working in a shared environment, which is closely related to the literature on collaborative communication (e.g. Two Body Problem (Jain et al. 2019)). CerealBar (Suhr et al. 2019), DialFRED (Gao et al. 2022) and TEACh (Padmakumar et al. 2022) introduce collaborative tasks where the human works as a disembodied issuer of instructions, possibly responding to robot's questions via explicit messages. In contrast, we consider the problem in which the AI assistant needs to consider both explicit messages in natural language and the implicit information in observed human actions. Further, we assume that instructions are not accurately describing the required information, but are generated based on a trade-off between informativeness and communication cost. Basically, we hope that robots can interpret ambiguous instructions without always needing clarification (asking questions). To this end, we focus on the HandMeThat (Wan, Mao, and Tenenbaum 2022) (HMT) benchmark, that calls for the ability to consider both explicit and implicit messages when following ambiguous instructions. The previous state-of-the-art work (Cao et al. 2024) on HMT performs iterated goal inference over the goal space in symbolic representation. However, it requires hand-crafted, pre-defined structures and extensive domain knowledge, which is not applicable in real-world scenarios.

Goal recognition. In our method, we hope to infer the human's intentions based on the observed historical actions, which is related to goal recognition problem (Lesh and Etzioni 1995; Baker, Tenenbaum, and Saxe 2007; Levesque 2011; Meneguzzi and Pereira 2021). Most of the works are based on the assumption of rationale that an agent should make (approximately) optimal decisions towards the goals every step (Dennett 1987; Gergely et al. 1995). Understanding human intentions in embodied environments has also been studied in many works; for example in Watch-and-Help (Puig et al. 2021) the AI must infer the human's goal from demonstrations, but no natural language is involved. Some recent works (Ying et al. 2024; Zhang et al. 2024) leverage LLMs to conduct goal inference based on the observed human actions or messages. In this work, we employ a small language model trained from scratch to undertake this part of the reasoning, since the aim is not to achieve precise goal recognition but to assist with the step-by-step social reasoning process.

Reasoning with intermediate steps. This work is also inspired by the research that uses intermediate steps to solve complex reasoning problems, including formal and mathematical reasoning and program synthesis (Roy, Vieira, and Roth 2015; Amini et al. 2019; Chiang and Chen 2019; Chen et al. 2020; Nye et al. 2021). Specifically, Nye et al. shows

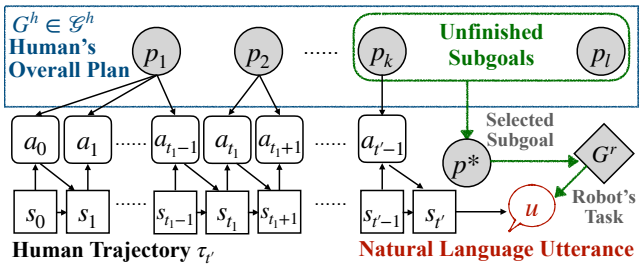

Figure 2: Problem formulation and proposed method. White nodes are observable, while grey nodes are unobservable. The robot is given the trajectory $\tau_{t'}$, a final state $s_{t'}$, and an utterance u. We propose to explicitly model human's intentions by modeling the human's overall plan $G^h \in \mathcal{G}^h$ as a set of predicates p_k. We further assume that human selects a subgoal p^* that needs help, and then specifies a robot's task G^r, which is the underlying intention when saying u.

that step-wise prediction method performs better than directly predicting the final outputs in program synthesis when prompting LLMs. Chain-of-Thought (CoT) (Wei et al. 2022) thoroughly explores how generating intermediate reasoning steps improves the performances of prompting LLMs to deal with complex reasoning tasks. In this paper, we show that social reasoning tasks benefit from the same approaches, and demonstrate that inferring human intentions is a critical component of successful human-robot collaboration.

FISER: Follow Instructions with Social and Embodied Reasoning

Problem Formulation. A human-robot Markov Decision Process is described as a tuple $\langle \mathcal{S}, \mathcal{A}^{h,r}, \mathcal{T}, \mathcal{U}, R^r, \gamma, T \rangle$. $s \in \mathcal{S}$ are object-oriented states including the locations, status and type of each object and agent. $\mathcal{A}^{h,r}$ is the joint action space with $\mathcal{A}^h, \mathcal{A}^r$ being the sets of actions available to the human and the robot, respectively. $\mathcal{T} : \mathcal{S} \times \mathcal{A}^{h,r} \times \mathcal{S} \to \{0, 1\}$ is the transition function where $\mathcal{T}(s, a^{h,r}, s') = 1$ if and only if taking actions $a^{h,r}$ at state s gives s' as the next state. $\mathcal{U}$ is a set of instructions that the human can give to the robot. $R^r : \mathcal{S} \to \mathbb{R}$ is a reward function for the robot, γ is the discount factor, and T is the horizon. Throughout the paper, we consider a scenario with only a single round of instruction following for the robot. In each episode, starting from an initial state s_0, the human begins working in the environment, and the robot is waiting. Human stops at a time step $t' \leq T$, leading to a trajectory $\tau_{t'} = (s_0, a_0^h, s_1, a_1^h, \ldots, s_{t'-1}, a_{t'-1}^h)$ and a final state $s_{t'}$. Then the human produces a natural language instruction $u \in \mathcal{U}$ that asks the robot for help. Given $\tau_{t'}, s_{t'}$ and u, the robot needs to interact with the environment by taking a sequence of actions $\{a_t^r\}_{t \geq t'}$ to maximize its discounted rewards $\sum_{t=t'}^{T} [\gamma^{t-t'} R^r(s_{t+1})]$.

Modeling the Human's Intentions

A straightforward solution to the human-robot MDP may treat $\tau_{t'}$ and u as additional state information. However, in reality, $\tau_{t'}$, u, and R^r have important correlations: when the

human is taking actions and producing instructions, their behavior can be modeled as optimizing for an internal reward function $R^h : \mathcal{S} \to \mathbb{R}$, which is not revealed to the robot. Our insight into this broad problem class is to leverage this causal relation between human's behavior and instruction by explicitly modeling hidden, unobserved variables representing human goals and intentions, so that we can make better use of the human trajectory to disambiguate the instruction (recognize the robot's task assigned by the human).

We start by assuming the reward functions R^h can be parameterized by a set of possible goals $\mathcal{G}^h$. The human's goal $G^h \in \mathcal{G}^h$ is sampled from an underlying distribution over $\mathcal{G}^h$ at the beginning of an episode, and is fixed across the horizon. However, it is not revealed to the robot directly. The goal in $\mathcal{G}^h$ is usually global and complex, such as "organize the bedroom." We assume the human trajectory τ_t was rational, produced to maximize the reward $R^h(\cdot \mid G^h)$. Based on G^h and the current progress $\tau_{t'}$, the human then selects a subgoal $p^* \in G^h$ (a part of the human's overall plan) that needs help, such as "having all books put in the box," and then specifies a task G^r for the robot (e.g., asking the robot to hand over a specific book). The instruction u is generated based on G^r. The relations between these variables are illustrated in Fig. 2.

Although we do not put specific assumptions over the structure of goals in $\mathcal{G}^h$ and how τ_t is generated, we illustrate them with a simplified example in Fig. 2. We define P as a set of predicates where each $p \in P$ is a classifier over states (to say whether the predicate is satisfied or not). For example, one predicate can be written as $\langle \exists y, \text{box}(y), \forall x, \text{book}(x) \Rightarrow \text{in}(x, y) \rangle$, which describes putting all apples in a box. Now we assume that the human goal $G^h = \{p_1, p_2, \cdots, p_l\}$ is a set of predicates. The human chooses to work on predicates one by one and has been working on all p_k's ($1 \le k \le l$) before stopping at time t', and then the subgoal p^* is chosen from the set of remaining predicates: $p^* \in \{p_k, \ldots, p_l\}$. Next, the human specifies a robot's task G^r such that the robot actions will result in a state s' where $R^h(s' \mid \{p^*\}) > 0$ (i.e., G^r is a useful step towards p^*, a remaining subgoal to accomplish). Note that neither p^* nor G^r is accessible to the robot, since the robot can only get access to the natural language instruction u. For example, *"could you pass that from the sofa"* could be an utterance for $G^r = \langle \text{human-holding(book\#0)} \rangle$ and $p^* = \langle \exists y, \text{box}(y), \forall x, \text{book}(x) \Rightarrow \text{in}(x, y) \rangle$.

Step-wise Reasoning over Human Intentions

Our model, FISER, builds on top of the factorized human-robot MDP formulation above. We formulate the problem into the social and embodied reasoning phases.

Social Reasoning: Robot's Task Recognition The robot needs to disambiguate the natural language instruction u into an understandable and executable task within its own goal space based on the observation of current state $s_{t'}$ and the historical trajectory $\tau_{t'}$. Therefore, we hope to estimate a function TR, such that $\text{TR}(s_{t'}, \tau_{t'}, u) \to G^r$.

Social Reasoning: Human's Plan Recognition. We further propose a variant that explicitly estimates human's underlying overall plan G^h based on $\tau_{t'}$ and replaces that trajectory by the predicted goal when doing instruction disambiguation. However, since recognizing the full plan is usually intractable, we opt to also take in u and $s_{t'}$, and directly predict the predicate $p^* \in G^h$ (subgoal) that the human wants the robot to help with. Therefore, we learn two functions PR and TR, such that $\text{PR}(s_{t'}, \tau_{t'}, u) \to p^*$ and $\text{TR}(s_{t'}, p^*, u) \to G^r$.

Embodied Reasoning: Grounded Planning. Once the robot goal G^r is obtained, the problem is reduced to a pure grounding and planning task. We can replace the ambiguous natural language instruction u by the accurately expressed robot goal G^r. The final grounded planning function GP should satisfy that $\text{GP}(s_{t'}, \tau_{t'}, G^r) \to \{a_t^r\}_{t \ge t'}$, which is basically learning a typical goal-conditioned robot policy $\pi(a | s_{y'}, \tau_{t'}, G^r)$*.

Since the functions TR and PR involve natural language inputs, language models are required for these two modules. For GP, we can either implement planning algorithms or use neural networks, since all inputs can be symbolic.

Transformer-based Model Implementation

We implement a Transformer-based model following our framework, illustrated in Fig. 3. We assume all inputs are rendered in texts, and the model needs to predict action strings. Since small-scale language models cannot process excessively long inputs, we divide the information into four parts, including world state description ($s_{t'}$), human's trajectory ($\tau_{t'}$), language instruction (u), and the model's past outputs.

World state description. We consider an object-centric representation for the world state. Specifically, the world state is described as a sequence of object tokens. For each object in the world, we fuse the information of its category (object type and genre), attributes (e.g., *size, color, is-open*), and spatial relation (*inside* or *on top of* another object) into one single embedding, which we called as an object token. The human and the robot are also treated as two special "objects".

Human's trajectory. A straightforward way to represent human's trajectory is to directly use a paragraph of texts to describe action sequence, e.g., "*the human picks up book#1 from the table*". To better align it with the world state, we replace the embeddings for object names ("book#1") by the object tokens that we obtained in the world state description.

Language instruction. The natural language instruction sentence is tokenized and then directly turned into embeddings.

Model's past outputs. Every time the agent takes a step, the environment returns a sentence describing the effect of its action, i.e., the update in observations. In each episode, such sentences for past steps are concatenated and served as an extra input, e.g., "*... [SEP] pick up book 0 [SEP] You pick up the book (book 0) from the table...*" The concatenated result is also tokenized and then turned into embeddings. This information is necessary because the model needs to know what it has done and whether the world state is changed.

Model Architecture Overview

The proposed model consists of $3N$ ($N = 3$) encoder layers that are used to update the representation over all four parts of the inputs, in order to predict human intentions or robot's actions. Specifically, we use the embeddings at Layer $2N$ to predict the robot's task G^r (social reasoning phase). Then, we

*$\tau_{t'}$ should be p^* instead if the PR stage is included.

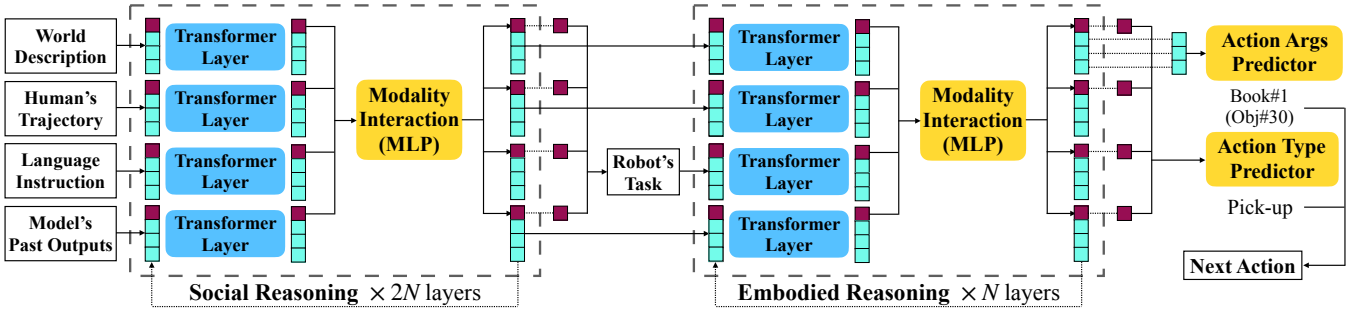

Figure 3: The Transformer-based model has four parts of inputs, which are passed separately into different Transformer Encoder Layers, and interact with each other through a Modality Interaction module after each Transformer layer. The first $2N$ layers form the social reasoning phase and the last N layers form the embodied reasoning phase. The embeddings at Layer $2N$ are used for recognizing robot's task, and the last layer embeddings are used for predicting actions.

replace the instruction input for Layer $2N+1$ by the predicted G^r and then use the last layer embeddings to predict robot's actions. If the Human's Plan Recognition stage is further included, we use the embeddings at Layer N to predict the selected human subgoal p^* and then replace the trajectory input for Layer $N+1$ by the predicted p^*.

The model is trained in either a multi-staged (MS) or an end-to-end (E2E) manner. The E2E models are trained to directly output robot's actions, but an auxiliary loss is applied over their intermediate predictions of robot's task (and human's plan). The MS models, however, disentangle the social reasoning (functions TR and PR) from embodied reasoning (function GP), and train them as two separate modules. The latter module is trained with the ground-truth p^* or G^r, but is evaluated using the predictions from the former module.

Encoder Layers

Each encoder layer is composed of four Transformer layers and a Modality Interaction module.

Transformer Layers. We use four separate Transformer encoder-only layers to process the inputs. We remove the positional encoding for the world state description, because we do not expect the model to learn an ordering of objects.

Modality Interaction. In order to fuse the information from the four parts of inputs, we design a modality interaction module (an MLP) within each layer, following the architecture proposed by GreaseLM (Zhang et al. 2021). We reserve a special "interaction" token at the front of each part of inputs. These tokens are expected to gather respective information in the Transformer layers and then interact with each other through this MLP. The updated special tokens will then replace the original first token in each part of inputs.

Prediction Layers

Now we introduce our predictors for intermediate reasoning steps and the final actions. All the predictions are trained with cross entropy loss over corresponding supervisions.

Human's Plan Recognition. We assume that there is a vocabulary of concepts that allow us to represent human goals as first order logic predicates (e.g., $\langle \exists y, \text{box}(y), \forall x, \text{book}(x) \Rightarrow \text{in}(x, y) \rangle$). While such logical predicates could be more complex, here we assume human plans follow a simpler form: a Q(uantifier), a S(ubjective), a V(erb), and an O(bjective) (e.g., $\langle$for-all, book, inside, box$\rangle$). Therefore, the human's plan recognition module of our model needs to predict a tuple of four tokens. The prediction is conditioned on the embeddings of the four inputs. Specifically, we calculate the log-likelihood over all tuples as follows, where the predictions of the Subjective and Objective are conditioned on the predicted values of Quantifier and Verb:

$$\log \Pr[\text{Q, S, V, O}] \approx \log \Pr[\text{Q}] + \log \Pr[\text{V}] + \log \Pr[\text{S}|\text{Q, V}] + \log \Pr[\text{O}|\text{Q, V}],$$

Robot's Task Recognition. The main target of the entire social reasoning phase is to predict the task assigned to the robot, such as a specific object to manipulate. Note that in this step the model needs to specify a concrete object in the world, while the Subjective and Objective predictions in previous plan recognition step are object types.

Action Prediction. The model needs to output the next action in each step, which we assume to be a triple consisting of action type and one or two arguments, (e.g., move-to(sofa), put-into(book#0, box#1)). The action type prediction is conditioned on the embeddings of the four inputs, while the arguments are further conditioned on the predicted action type. Specifically, we calculate the log-likelihood over all actions (no matter applicable or not).

$$\log \Pr[\text{Action,Arg1, Arg2}] \approx \log \Pr[\text{Action}] + \log \Pr[\text{Arg1}|\text{Action}] + \log \Pr[\text{Arg2}|\text{Action}].$$

We assume the model has access to all applicable actions at each step, so we take the maximum over all applicable triples to get the final prediction.

Experiments

We evaluate our framework by training a Transformer-based model from scratch for the challenging HandMeThat benchmark (Wan, Mao, and Tenenbaum 2022), and then compare them with multiple competitive baselines, including the state-of-the-art prior work on HMT, and the CoT prompting on the largest available pre-trained language models.

HandMeThat Environment

We evaluate our models over the HandMeThat (version 2) dataset (Wan, Mao, and Tenenbaum 2022). It introduces a household ambiguous instruction following task rendered in text. HandMeThat instructions are split into four difficulty levels, and the gaps between levels correspond to different challenges. The instruction in a Level 1 task has no ambiguity—it is a pure planning task. A Level 2 task requires social reasoning where a robot can successfully accomplish the task if it can also infer the goal from the human trajectory. On Level 3, the robot needs to further consider pragmatic reasoning in language use. For example, if there are multiple books everywhere in the room and only one coat on the sofa, where both the book and the coat are helpful to the human's goal, the human may not refer to the coat by saying "*Could you pass that from the sofa*" since "*from the sofa*" is a redundant specification in that case. The final Level 4 contains tasks with inherent ambiguities that cannot be resolved with the existing information, but can potentially be resolved with a strong prior over what human is likely to do. We evaluate all the models on their success rates to achieve the robot's goal. Note that the original evaluation metric in HandMeThat additionally considers the number of robot's steps. An agent can trivially improve success rate with increased steps, by simply searching in a trial-and-error fashion. We believe that enumeration over objects is not realistic in the real world, so we restrict our experiments to one trial (within 4 or 5 steps).

Model Details

Baseline models. We compare our results to human performance on the task (Human), a hand-coded baseline (Heuristic) which have access to ground-truth symbolic state information, and a neural network baseline (Seq2Seq (Sutskever, Vinyals, and Le 2014)) introduced in the HandMeThat paper. The existing SOTA work (Cao et al. 2024) was implemented based on the original HandMeThat (version 1) dataset. Therefore, we report the results of our FISER models over original data points that lie in the version 2 domain. Further details of this comparison are provided in the supplementary material.

Transformer-based model. Following the proposed model architecture, we implement a set of Transformer-based models. We compare the implementations with no intermediate supervision (Transformer), with Robot's Task Recognition only (Transformer+FISER), and with Human's Plan Recognition in addition (Transformer+FISER+PR). Our models are trained from scratch because existing small-scale pre-trained models cannot handle the excessive token lengths of HandMeThat data inputs. We compare two ways of training the Transformer-based model using FISER framework, end-to-end (E2E) or multi-staged (MS), as explained in the overview of model architecture. This comparison aims to provide insights for whether to block the gradient flow from embodied reasoning back to the social reasoning module. We further report the accuracy of the intermediate prediction steps for these Transformer-based models, including QSVO (simplified human subgoal), and Obj (concrete object to be manipulated in expert demonstration, i.e., the robot's task). The number of parameters in an E2E model is 5.1M; the number of parameters in an MS model is 4.7M.

Prompted GPT-4. We design prompting methods for GPT-4 Turbo over HMT tasks. The vanilla implementation (GPT-4) simply provides all inputs to the model and requests it to output actions. We first conduct prompt engineering (GPT-4+PE) to incorporate some domain-specific knowledge and help to parse the complex inputs. Then we implement FISER framework by applying CoT prompting to explicitly predict the same intermediate data (human's plans and specified robot's tasks) that we use for Transformer-based models step-by-step, which similarly gives two models (i.e., GPT-4+FISER and GPT-4+FISER+PR). To be more specific, in both social reasoning steps, we prompt GPT-4 to do step-by-step reasoning except that we ask different questions. Human's Plan Recognition asks about the human's higher-level goal, while Robot's Task Recognition asks about the intended meaning of an ambiguous instruction.

Results

We evaluate all the models over the HandMeThat (version 2) dataset in the fully observable setting. Overall, our best-performing Transformer+FISER model achieves a 64.5% success rate on average, achieving the state-of-the-art on the HandMeThat benchmark. The main results are presented in Table 1. Now we discuss the following hypotheses.

H1: Explicitly modeling human intentions works better than directly predicting actions.

(a) Separating the social and embodied reasoning steps by explicitly recognizing the robot's task is beneficial.

For both prompted GPT-4 Turbo and Transformer-based models, explicitly predicting the robot's task significantly improves the success rates across all difficulty levels, which supports our hypothesis that separating the social and embodied reasoning steps is beneficial in these complex reasoning tasks. The comparison between two different training schemes of our Transformer-based models are presented in Table 2. Results show that training in a multi-staged manner works better than end-to-end in our tasks. It may imply that the low-level grounded planning (embodied reasoning) is requiring a sufficiently different task representation from inferring human's internal goals (social reasoning), that allowing gradients from the embodied reasoning module to flow into the social reasoning module actually hurts performance. It is a further support on empirical side that we should make explicit inferences about human intentions as intermediate reasoning steps.

(b) Explicitly recognizing the human's plan further helps with the social reasoning stage.

When we further include the Human's Plan Recognition (PR) stage, we find that it only helps for the most ambiguous cases (like in Level 4). For GPT-4 Turbo, adding PR is showing approximately the same performances as normal FISER method. We hypothesize that pre-trained LLMs are not good at leveraging hierarchical predictions to improve on this task. For Transformer-based models, introducing PR gives better performance on Level 4, but is harmful to the simplest Level 1. We attribute the poor performance in Level 1 to the fact that such simplest tasks do not require knowing the humans' high-level goal. Therefore, forcing the model to predict this information reduces model's capacity to focus

Model	Baseline Models			GPT-4 Turbo				Transformer-based Models			On HMT Version 1	
	Human	Heur.	Seq2Seq	Vanilla	+PE	+FISER	+PR	Vanilla	+FISER	+PR	Cao et al.	FISER
Level 1	100.0	100.0	30.4	72.0	82.0	80.0	77.0	77.7±1.6	**89.0**±1.5	72.0±1.5	27.7±0.3	89.7±0.4
Level 2	80.0	64.0	28.8	16.0	25.0	36.0	34.0	55.3±0.4	**74.0**±0.3	**74.0**±0.9	24.8±0.4	63.0±0.3
Level 3	40.0	39.0	12.8	5.0	13.0	18.0	17.0	36.3±1.0	**52.3**±2.3	**52.3**±1.0	21.0±0.1	28.3±2.5
Level 4	30.0	29.0	14.8	9.0	9.0	17.0	20.0	38.3±1.4	42.7±0.2	**51.0**±1.2	21.7±0.2	40.7±1.9

Table 1: Success rate (%) of models over HMT in the fully observable setting. The results for Transformer-based models are the mean and standard error values over three runs. *Cao et al. (2024) is evaluating version 1 of HandMeThat dataset, and thus we provide the results of Transformer+FISER model over a subset of version 1 for a fair comparison. Overall, FISER improves the performance across all levels compared to the vanilla Transformer. While applying FISER and PR to GPT-4 improves its performance, overall GPT-4 cannot perform well on these tasks even with very careful prompting, achieving less than half the success rate of our model for ambiguous instructions in levels 2-4.

Model	Level 1	Level 2	Level 3	Level 4
End-to-End	77.7±1.6 (N/A, N/A)	55.3±0.4 (N/A, N/A)	36.3±1.0 (N/A, N/A)	38.3±1.4 (N/A, N/A)
End-to-End+FISER	74.3±0.2 (N/A, 87.8)	73.7±0.5 (N/A, 80.9)	47.3±2.1 (N/A, 73.1)	41.3±1.9 (N/A, 64.7)
End-to-End+FISER+PR	61.0±1.5 (73.2, 81.3)	66.7±1.1 (64.9, 81.0)	47.0±1.2 (59.0, 73.4)	42.0±0.3 (55.0, 66.5)
Multi-Staged+FISER	**89.0**±1.5 (N/A, 93.4)	**74.0**±0.3 (N/A, 82.3)	**52.3**±2.3 (N/A, 76.5)	42.7±0.2 (N/A, 68.1)
Multi-Staged+FISER+PR	72.0±1.5 (74.4, 82.4)	**74.0**±0.9 (71.9, 82.3)	**52.3**±1.0 (67.1, 75.6)	**51.0**±1.2 (65.3, 71.2)

Table 2: Comparison between End-to-End and Multi-staged training of Transformer-based models over HMT in the fully observable setting. All models are evaluated by the success rate (%). The prediction accuracy (%) of QSVO and Obj (intermediate outputs) are presented in parentheses. The results are the mean values over three runs, and the standard error values for success rates are provided. Overall, training in a multi-staged manner works better than end-to-end in our tasks, implying that fully separating social from embodied reasoning provides the best performance.

on planning for low-level actions. On the other hand, the improved performance on Level 4 shows that explicit human's plan recognition helps to better learn priors over human intentions. Even on these intrinsically ambiguous tasks, the model can leverage the strong prior to take helpful actions.

H2: Pre-trained LLMs, despite having access to commonsense knowledge, do not adequately perform the complex social and embodied reasoning in this task. Incorporation of domain-specific knowledge through CoT can help.

Results show that training much smaller, more efficient Transformer-based models from scratch is exhibiting about 70% increased performance than prompting state-of-the-art pre-trained LLMs. GPT-4 Turbo's results on Level 1 show it has the capability to do some level of embodied reasoning when given explicit tasks. However, the performance drop on subsequent levels indicates that the required knowledge to solve HandMeThat tasks is not fully covered by commonsense knowledge in pre-trained LLMs. With well-designed prompt engineering (PE) that contains some domain knowledge (e.g., goal space and few-shot examples), GPT-4 Turbo improves significantly across all difficulty levels. However, even with careful CoT prompts and few-shot examples, GPT-4 Turbo is far from small-scale Transformer models across all levels. As a qualitative analysis, its failure modes on Level 2 consists of 36% planning failure (hallucination or invalid actions), 31% incorrect intention (common-sense reasoning that are not aligned with the ground-truth), and 23% redundant behavior (giving human an object that is already at its target position). We believe this is because the prompting methods alone cannot provide the model with the type of social and embodied reasoning needed to solve this task. The information learned from language datasets collected online may also be significantly different from that required for this household assistance task. Training a small-scale model, however, can solve the problem more efficiently and reliably.

We conducted an additional experiment in the extended version to see if the performance of the pre-trained LLMs could be improved. By filtering out a proportion of objects in the world that are irrelevant to the human's task, which assumes access to the ground-truth human goals. We observe that LLM's performance relies on a very large proportion of objects being filtered out, which provides further insight that LLMs cannot effectively select information and focus on relevant objects, which is required in embodied reasoning.

Conclusion

We study the challenging HandMeThat benchmark, comprising ambiguous instruction following tasks requiring sophisticated embodied and social reasoning. We find that existing approaches for training models end-to-end, or for prompting powerful pre-trained LLMs, are both insufficient to solve these tasks. We hypothesized that performance could be improved by building a model that explicitly performs social reasoning to infer the human's intentions from their prior actions in the environment. Our results provide evidence for this hypothesis, and show that our approach, Follow Instructions with Social and Embodied Reasoning (FISER), enhances performance over the most competitive prompting baselines by 70%, setting the new state-of-the-art for HandMeThat.

Acknowledgments

We express our gratitude to the anonymous reviewers for their valuable comments and suggestions. We also thank our friends and colleagues for their insightful feedback. This research was supported by the Cooperative AI Foundation and the UW + Amazon Science Gift fund. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of our sponsors.

References

Ahn, M.; Brohan, A.; Brown, N.; Chebotar, Y.; Cortes, O.; David, B.; Finn, C.; Fu, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; Ho, D.; Hsu, J.; Ibarz, J.; Ichter, B.; Irpan, A.; Jang, E.; Ruano, R. J.; Jeffrey, K.; Jesmonth, S.; Joshi, N. J.; Julian, R.; Kalashnikov, D.; Kuang, Y.; Lee, K.-H.; Levine, S.; Lu, Y.; Luu, L.; Parada, C.; Pastor, P.; Quiambao, J.; Rao, K.; Rettinghouse, J.; Reyes, D.; Sermanet, P.; Sievers, N.; Tan, C.; Toshev, A.; Vanhoucke, V.; Xia, F.; Xiao, T.; Xu, P.; Xu, S.; Yan, M.; and Zeng, A. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. arXiv:2204.01691.

Amini, A.; Gabriel, S.; Lin, S.; Koncel-Kedziorski, R.; Choi, Y.; and Hajishirzi, H. 2019. MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms. In *NAACL-HLT*, 2357–2367.

Baker, C. L.; Tenenbaum, J. B.; and Saxe, R. 2007. Goal Inference as Inverse Planning. *CogSci*.

Blukis, V.; Paxton, C.; Fox, D.; Garg, A.; and Artzi, Y. 2021. A Persistent Spatial Semantic Representation for High-level Natural Language Instruction Execution. In *CoRL*, volume 164, 706–717.

Cao, C.; Fu, Y.; Xu, S.; Zhang, R.; and Li, S. 2024. Enhancing Human-AI Collaboration Through Logic-Guided Reasoning. In *ICLR*.

Chen, X.; Liang, C.; Yu, A. W.; Zhou, D.; Song, D.; and Le, Q. V. 2020. Neural Symbolic Reader: Scalable Integration of Distributed and Symbolic Representations for Reading Comprehension. In *ICLR*.

Chiang, T.-R.; and Chen, Y.-N. 2019. Semantically-Aligned Equation Generation for Solving and Reasoning Math Word Problems. In Burstein, J.; Doran, C.; and Solorio, T., eds., *NAACL-HLT*, 2656–2668.

Clark, H. H. 1996. *Using language*. Cambridge university press.

Dennett, D. C. 1987. *The intentional stance*. MIT press.

Gao, X.; Gao, Q.; Gong, R.; Lin, K.; Thattai, G.; and Sukhatme, G. S. 2022. DialFRED: Dialogue-Enabled Agents for Embodied Instruction Following. *RA-L*, 7: 10049–10056.

Gergely, G.; Nádasdy, Z.; Csibra, G.; and Bíró, S. 1995. Taking the intentional stance at 12 months of age. *Cognition*, 56(2): 165–193.

Grice, H. P. 1975. Logic and Conversation. *Syntax and Semantics*, 3: 41–58.

Jain, U.; Weihs, L.; Kolve, E.; Rastegari, M.; Lazebnik, S.; Farhadi, A.; Schwing, A. G.; and Kembhavi, A. 2019. Two Body Problem: Collaborative Visual Task Completion. In *CVPR*.

Kojima, N.; Suhr, A.; and Artzi, Y. 2021. Continual Learning for Grounded Instruction Generation by Observing Human Following Behavior. *TACL*, 9.

Lesh, N.; and Etzioni, O. 1995. A Sound and Fast Goal Recognizer. In *IJCAI*.

Levesque, R. J. R. 2011. *Social Reasoning*, 2808–2808. Springer New York. ISBN 978-1-4419-1695-2.

Meneguzzi, F.; and Pereira, R. 2021. A Survey on Goal Recognition as Planning. In *IJCAI*.

Min, S. Y.; Puig, X.; Chaplot, D. S.; Yang, T.-Y.; Rai, A.; Parashar, P.; Salakhutdinov, R.; Bisk, Y.; and Mottaghi, R. 2024. Situated Instruction Following. In *ECCV*.

Nair, S.; Rajeswaran, A.; Kumar, V.; Finn, C.; and Gupta, A. 2022. R3M: A Universal Visual Representation for Robot Manipulation. In *CoRL*, volume 205, 892–909.

Nye, M.; Andreassen, A. J.; Gur-Ari, G.; Michalewski, H.; Austin, J.; Bieber, D.; Dohan, D.; Lewkowycz, A.; Bosma, M.; Luan, D.; Sutton, C.; and Odena, A. 2021. Show Your Work: Scratchpads for Intermediate Computation with Language Models. arXiv:2112.00114.

Padmakumar, A.; Thomason, J.; Shrivastava, A.; Lange, P.; Narayan-Chen, A.; Gella, S.; Piramithu, R.; Tur, G.; and Hakkani-Tür, D. Z. 2022. TEACh: Task-driven Embodied Agents that Chat. In *AAAI*.

Puig, X.; Shu, T.; Li, S.; Wang, Z.; Liao, Y.-H.; Tenenbaum, J. B.; Fidler, S.; and Torralba, A. 2021. Watch-And-Help: A Challenge for Social Perception and Human-{AI} Collaboration. In *ICLR*.

Roy, S.; Vieira, T.; and Roth, D. 2015. Reasoning about Quantities in Natural Language. In *TACL*, 1–13.

Shridhar, M.; Thomason, J.; Gordon, D.; Bisk, Y.; Han, W.; Mottaghi, R.; Zettlemoyer, L.; and Fox, D. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *CVPR*.

Siskind, J. M. 1994. Grounding language in perception. *JAIR*, 8: 371–391.

Sperber, D.; and Wilson, D. 1986. *Relevance: Communication and cognition*. Citeseer.

Suglia, A.; Gao, Q.; Thomason, J.; Thattai, G.; and Sukhatme, G. 2021. Embodied BERT: A Transformer Model for Embodied, Language-guided Visual Task Completion. arXiv:2108.04927.

Suhr, A.; Yan, C.; Schluger, J.; Yu, S.; Khader, H.; Mouallem, M.; Zhang, I.; and Artzi, Y. 2019. Executing Instructions in Situated Collaborative Interactions. In *EMNLP-IJCNLP*.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to Sequence Learning with Neural Networks. In *NeurIPS*.

Wan, Y.; Mao, J.; and Tenenbaum, J. 2022. HandMeThat: Human-Robot Communication in Physical and Social Environments. In *NeurIPS Datasets and Benchmarks Track*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*.

Winograd, T. 1972. Understanding natural language. *Cognitive Psychology*, 3(1): 1–191.

Ying, L.; Jha, K.; Aarya, S.; Tenenbaum, J. B.; Torralba, A.; and Shu, T. 2024. GOMA: Proactive Embodied Cooperative Communication via Goal-Oriented Mental Alignment. arXiv:2403.11075.

Zellers, R.; Holtzman, A.; Peters, M.; Mottaghi, R.; Kembhavi, A.; Farhadi, A.; and Choi, Y. 2021. PIGLeT: Language Grounding Through Neuro-Symbolic Interaction in a 3D World. In *ACL*, 2040–2050.

Zhang, H.; Wang, Z.; Lyu, Q.; Zhang, Z.; Chen, S.; Shu, T.; Du, Y.; and Gan, C. 2024. COMBO: Compositional World Models for Embodied Multi-Agent Cooperation. arXiv:2404.10775.

Zhang, X.; Bosselut, A.; Yasunaga, M.; Ren, H.; Liang, P.; Manning, C. D.; and Leskovec, J. 2021. GreaseLM: Graph REASoning Enhanced Language Models. In *International Conference on Learning Representations*.

Zhi-Xuan, T.; Ying, L.; Mansinghka, V.; and Tenenbaum, J. B. 2024. Pragmatic Instruction Following and Goal Assistance via Cooperative Language-Guided Inverse Planning. In *AAMAS*.