# Intuition Estimation and Knowledge-Based Planning for Human-AI Collaboration

Zihan Ding[1*], Jinyu Chen[1*], Si Liu[1†], and Shifeng Zhang[2]

[1] Institute of Artificial Intelligence, Beihang University
[2] Sangfor Technologies Inc.

**Abstract.** Humans possess an innate ability to infer others' intentions from ambiguous utterances based on the observation of contextual cues and past actions. Conversely, machines typically necessitate explicit instructions, thereby increasing the temporal cost of human-AI interaction. To mitigate this, we propose the Intuition Estimation and Knowledge-Based Planning (IEKP) method, which augments human-AI collaboration under ambiguous directives. IEKP encompasses three principal components: 1) Associative Reasoning based Goal Recognition (`ARGoal`) utilizes large language model to form an initial estimation of human goals and refines this estimation through associative mechanisms; 2) Finite State Machine Guided Decision Pruning (`FDPrune`) constructs state machines based on task types, pruning illegitimate action outputs to enhance the robustness of language models in long-term decision processes; 3) Knowledge-Enhanced Searching System (`K-Search`) leverages co-occurrence relationships between objects and environments to improve the agent's efficiency in environmental searches. Our approach markedly enhances performance on the HandMeThat task, increasing the success rate by 65.42% and the average score by 88.03 compared to previous state-of-the-art methods, even surpassing human performance. This underscores the efficacy of IEKP in advancing human-AI collaboration through superior comprehension and execution of under-specified instructions.

**Keywords:** Human-AI Collaboration · Intuition Estimation · Knowledge-Based Planning.

## 1 Introduction

Humans possess an innate ability to infer others' intentions [10]. During human-to-human communication and collaboration, unnecessary information is often omitted. For instance, when one person places a fruit on the counter and asks another to "hand me that" [39], it is easy to infer that the desired object is a fruit knife and he wants to cut the fruit up. This ability to understand intentions stems from an understanding of the surrounding environment and historical actions. However, when communicating with machines, explicit instructions are often required, increasing the time cost of human-AI interaction. Therefore, future

*Equal contribution.
†Corresponding author.

assistant agents should be able to observe human behavior and assist humans based on under-specified human utterances, which is of immense importance for home care services and simplify the process of usage.

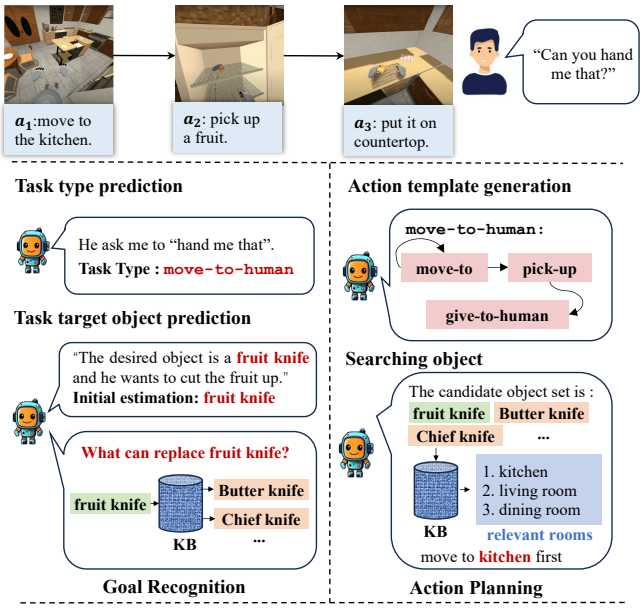


**Fig. 1.** In IEKP, there are two main steps: goal recognition and action planning. During goal recognition, the task type is predicted and the initial estimation of task target objects is expanded using substitution knowledge. In the action planning process, an action template is generated based on the task type, and the search room sequence is determined using co-occurrence knowledge.

In robotic planning for executing human ambiguous instructions in household scenarios, three key aspects must be addressed: (1) identifying the target objects of human interest and the appropriate actions to perform on these target objects, (2) locating suitable positions to complete the requested tasks (3) and generating executable action planns. By progressively resolving these issues, the robot can effectively disambiguate vague human instructions.

As shown in Fig. 1, referring to the observations mentioned above, we propose the *Intuition Estimation and Knowledge-Based Planning* (IEKP) mechanism, which leverages common-sense knowledge and the large language model (LLM) to execute human ambiguous instructions in household scenarios. IEKP has three components, namely *Associative Reasoning based Intention Prediction* (`ARGoal`), *Knowledge-Enhanced Searching System* (`K-Search`) and *Finite State Machine Guided Decision Pruning* (`FDPrune`). Firstly, `ARGoal` will recognize the human's target objects of human interest and the appropriate actions,

*i.e.* the task type with LLM. Additionally, we observe that objects of similar usage can be substituted for some goals, such as bottles and cups. Therefore, we have constructed the object of interest substitution prior, and use this prior to infer other possible substitutes to guide subsequent planning. Secondly, the process of searching for objects plays an important role to house-holding tasks. The `K-Search` infers the order of exploration based on the environment-object co-occurrence prior. The `K-search` will generate a room search sequence based on the co-occurrence probability of rooms and objects to find the subtile room to fulfill the task, which enhance the efficiency of environmental searching especially for more ambiguous instructions. Thirdly, `FDPrune` prunes the action sequence generated from LLM to an executable action plan. Since LLMs often suffer from hallucinations [18] when generating action plans, leading to unexecutable or repeating action sequences. `FDPrune` constructs a decision-making finite state machine based on the task type to regulate the actions generated by LLMs, thereby enhancing the effectiveness of the action plan.

With IEKP, we achieved significant performance improvements on HandMeThat [39], increasing the success rate by 65.42% compared to previous state-of-the-art methods at level 4 difficulty, and even outperforms human, proving the effectiveness of IEKP in human-AI collaboration tasks involving ambiguous instructions. To summarize, our contribution is three-fold: **i)** We propose an associative goal recognition method, `ARGoal`, to augment the understanding of ambiguous instructions. **ii)** We propose an efficient decision pruning method, `FDPrune`, to bolster the robustness of language models in predicting decisions. **iii)** We propose a co-occurrence based searching mechanism, `K-Search`, to enhance the efficiency on locating the required objects when the utterance is ambiguous.

## 2 Related Work

*Household Agent.* Enabling agents to understand natural language instructions is a critical capability for household robots, and methods like [12, 5, 42, 2, 20] rely on explicit language instructions for navigation. AlFworld [31] and ALFRED [30] are build upon AI-2THOR [19] and requires the agent to complete indoor manipulation tasks via language instructions. In these tasks, the objectives are typically clear and well-defined. However, in real-world human collaboration, instructions often contain implicit omissions to enhance communication efficiency. Understanding such ambiguous instructions is crucial for enhancing the efficiency of human-robot collaboration.

*Goal Recognition.* Cooperation based on ambiguous instructions is related to goal recognition: inferring the goals of other agents based on their historical actions [34, 3, 21, 28, 29, 13, 43, 23]. The prevailing assumption is the principle of rationality: agents are expected to make (approximately) optimal decisions to achieve their goals, given their beliefs [7, 14]. Understanding human intentions in embodied environments has been explored in various studies. In HandMeThat [39], the robot's objective is to ascertain the subgoal designated by

the human via utterance to help human to finish a task. Watch-and-Help [26] introduces a non-language goal inference task based on the VirtualHome environment [25]. CerealBar [33] proposes a benchmark for robotic instruction following in a collaborative environment. LIGHT [37] sets up a textual platform for understanding actions and emotions within natural language dialogues, focusing on background knowledge such as backstory and personality to interpret human's internal ideas. In this paper, we utilize LLM and co-occurrence knowledge to construct a set-based goal representation that accurately captures human ambiguous intentions.

*Knowledge Augmented Agent.* Knowledge has long been considered an effective means to assist human-AI interaction in various fields [11, 4, 32, 9, 8]. The field of augmented language models (ALMs) aims to mitigate the hallucinations observed in traditional LLMs by enhancing their reasoning abilities and enabling the use of external resources [6]. Recently, many studies [36, 22, 38, 27, 44] have been proposed to augment the reasoning capabilities of LLMs by integrating knowledge from external resources, including search engines, knowledge bases, and Wikipedia documents. LLMs can utilize these improvements either independently or by integrating them in a particular sequence to accomplish a designated task, ultimately leading to augmented capabilities [24, 41]. In this paper, we build co-occurrence knowledge from the household environment to help agent correct mistakes and refine search strategies.

## 3 Method

### 3.1 Problem Setup

In this paper, we investigate the task of human-AI collaboration based on ambiguous instructions. In this section, we will provide a formalized description for the task. The environment of HandmeThat [39] can be defined as $\langle \mathcal{S}, \mathcal{A}, \gamma, \mathcal{O} \rangle$, where $s \in \mathcal{S}$ represents the state space, $\mathcal{A}$ the action space, $\gamma$ the transition function, and $\mathcal{O}$ the observation function. An action $a = \langle \hat{a}, O_{\text{arg}} \rangle \in \mathcal{A}$, is represented as an action schema $\hat{a}$ along with the target objects $O_{\text{arg}}$, for instance, `move-to(bedroom)` or `robot-open(cabinet)`. When the agent in state $s$ takes the action $a$, the environment state deterministically transitions to $s' = \gamma(s, a)$, and the agent receives a new observation $o' = \mathcal{O}(s')$, reflecting the partial observation in indoor environments. In this task, each episode is characterized by the tuple $\langle \Omega, s_0, m, u \rangle$. The human agent performs $T$ actions starting from the initial state $s_{-T} \in \mathcal{S}$, generating a trajectory $\Omega = \{a_{-T}, a_{1-T}, \ldots, a_0\}$ and reach the final state $s_0$. Subsequently, the human sets a goal $m$ and specifies an utterance $u$ for the agent to reach, which may be ambiguous. The agent, upon observing $\langle \mathcal{O}(s_0), \Omega, u \rangle$, assists the human in achieving the goal $m$.

### 3.2 Overview

First, `ARGoal` involves goal recognition from the ambiguous instructions, which can be defined as:

$$\hat{m} = \arg\max_{m} p(m|\Omega, \mathcal{O}(s_0), u), \tag{1}$$

where $\hat{m}$ is the estimated human goal. We design a object-set-based representation for $\hat{m}$. Then, `FDPrune` will predict the task planning finite state machine (FSM) `tmpl`:

$$\texttt{tmpl} = \arg\max_{\texttt{tmpl}} p(\texttt{tmpl}|\Omega, \mathcal{O}(s_0), u, \hat{m}). \tag{2}$$

At state $s_t$, the `tmpl` defines a feasible action set $A_t$. The LLM only allows to predict actions $\hat{a}_t \in A_t$:

$$\hat{a}_t = \arg\max_{a \in A_t} p(a|\Omega, \mathcal{O}(s_0), u, \Omega_a), \tag{3}$$

where $\Omega_a$ is the historical action set of agent. When the agent predict the `move-to` action for searching objects, `K-Search` will leverage the location-object co-occurrence probability prior, combined with $\hat{m}$, to optimize the search sequence.

### 3.3 Associative Reasoning based Intention Prediction

When humans execute ambiguous instructions, such as HandMeThat [39], two aspects are primarily considered, the task type and the task target object. So the predicted $\hat{m}$ is defined as:

$$\hat{m} = \langle \mathcal{T}, M \rangle, \tag{4}$$

where $\mathcal{T}$ is the task type and $M$ is the candidate task target object set. The task type is often specified within the instruction, so we utilize LLMs to predict the $\mathcal{T}$ based on the given instruction:

$$\mathcal{T} = f_{\theta_0}(\Omega, \mathcal{O}, u), \tag{5}$$

where $f_{\theta_0}$ is an LLM with the action type predicting prompt.

For the task target object prediction, there are mainly two steps, *i.e.*, initial estimation and association. For the initial estimation process, where the agent need to infer a potentially helpful item based on the requester's past actions and the environment, such as guessing that a fruit knife might be suitable for him to cut an apple. We use an LLM to perfrom the initial estimation process. Given that LLMs have strong contextual understanding and are based on extensive pretraining, they possess a certain degree of intent inference capability. We adopt a set representation for task target objects $M_1 = \{\langle \lambda, p_\lambda \rangle\}$ for initial estimation. Here, $\lambda$ denotes the object involved in $m$, and $p_\lambda$ represents the probability

that the object meets $m$. We assume that the probabilities of different objects predicted by LLM ($p_{\theta_1}(\lambda)$, proportional to $p_\lambda$:

$$p_\lambda \propto p_{\theta_1}(\lambda) = p_{\theta_1}(\lambda|\Omega, \mathcal{O}(s_0), u). \tag{6}$$

We then select the top $k_1$ objects with the highest predicted $p_{\theta_1}(\lambda)$ to form the target object set, $M_1$ and use $p_{\theta_1}(\lambda)$ to replace $p_\lambda$. This inference might be inaccurate or the item might be absent in the environment.

For the associative ability, where humans consider other possible alternatives, such as whether a butter knife or a chef's knife might also be helpful for cut an apple. This associative ability enhances efficiency and success in executing instructions. Based on $M_1$, we employ an associative mechanism to expand other possible objects through substitution relationships. This substitution relationship assumes that different objects can achieve the same goal $m$ with a certain probability $p_r(\lambda_1|\lambda_2)$:

$$p_r(\lambda_1|\lambda_2) = \frac{\sum_m p(m) \cdot p(\lambda_1, \lambda_2|m)}{\sum_m p(m) \cdot p(\lambda_2|m)}, \tag{7}$$

where $p(m)$, $p(\lambda_1, \lambda_2|m)$ and $p(\lambda_2|m)$ can be estimated from the training set.

Therefore, the probability that $\lambda^*$ meets the goal requirement can be estimated via association with:

$$p_{\theta_1}(\lambda^*) = \sum_{\lambda \in M_1} p_{\theta_1}(\lambda) \cdot p_r(\lambda^*|\lambda). \tag{8}$$

Subsequently, we add the $\lambda^*$ with $p_\theta(\lambda^*)$ larger than $k_2$ to $M_1$ to obtain the final estimation of $\hat{m}$:

$$M = M_1 \cup \{\langle \lambda*, p_{\lambda*} \rangle | p_{\theta_1}(\lambda^*) > k_2\}, \tag{9}$$

where $k_2$ is a hyper-parameter. $M$ will serve as an estimation for the objective, assisting the LLM in generating action plans and inferring co-occurrence relationships for use in `K-search`.

### 3.4 Finite State Machine Guided Decision Pruning

For executing ambiguous instructions, generating an effective sequence of action decisions is also challenging. However, LLMs are not adept at generating standardized action code, often producing outputs with format or logic errors, making the generated action sequences difficult to execute robustly. Conversely, indoor tasks involve specific plan types, such as "bring me", "change state", and "move to", which can be standardized into finite-state machines. For example, the task type "bring to human" can be summarized as "[[`move-to`]+, [`open`]?, `pick-up`, `bring-to-human`]+" [44]. Using FSM allows the LLM to focus on understanding semantic information, which is its strength. This process involves two main steps: first, predicting the planning template based on the predicted

task type $\mathcal{T}$; second, using the action finite-state machine to assist the planning process. We first use a LLM to predict the action template required for the task:

$$\texttt{tmpl} = f_{\theta_2}(\Omega, \mathcal{O}(s_0), u, \mathcal{T}), \tag{10}$$

where $f_{\theta_2}$ is the LLM with the action template predicting prompt. `tmpl` can extract the set of actions $A_t$ schema available in this state:

$$A_t = \texttt{extract-action-set}(\texttt{tmpl}, s_t) \tag{11}$$

Thus, for an agent in state $s_t$, we select the action $\hat{a}_t$ with the highest probability as the next action type:

$$\hat{a}_t = \arg\max_{\hat{a} \in A_t} p_{\theta_3}(\hat{a}|\Omega, \mathcal{O}(s_t), u, \Omega_a), \tag{12}$$

where $\Omega_a$ represents the history of actions, and $\theta_3$ denotes the model parameters. We achieve this by masking out unreasonable vocabulary outputs. Based on $\hat{a}_t$, the LLM only needs to predict the subjects $O_{\text{arg}}$ of the action $\hat{a}_t$, such as the objects to be operated on or the rooms to move to. To further enhance the accuracy of the model's predictions, we leverage the LLM's contextual inference capability. We input the predicted subgoal representation $\hat{m}$ as context into the LLM to predict $O_{\text{arg}}$:

$$O_{\text{arg}} = \pi_{\theta_3}(\Omega, \mathcal{O}(s_t), u, \hat{m}, \hat{a}_t). \tag{13}$$

Using this method can effectively enhance the robustness of the LLM in the decision-making process based on action schemas.

### 3.5 Knowledge-Enhanced Search System

In the process of completing household tasks, regardless of the planning template, it is essential to locate target objects in the environment. Therefore, the ability to search within the environment becomes crucial. The agent needs to infer the possible locations of target objects based on instructions, and adjust its position based on observations. This makes the accuracy of predicting the action schema `move-to` and its $O_{\text{arg}}$ critical for the success rate and efficiency of the task. Our experiments reveal that language models struggle to correct their mistakes in real-time and adjust search strategies, often leading to repeated searches in the same area, which hinders effective task completion. To address this issue, we utilize co-occurrence knowledge to construct a heuristic search sequence that assists the LLM in `move-to` decisions. Since the distribution of objects in a household environment often follows specific co-occurrence rules [11]—such as pillows typically found in bedrooms and milk usually stored in refrigerators—effectively utilizing these co-occurrence relationships can improve the efficiency of searching for specific objects. The probability of object $\lambda$ appearing at location $l$ is $p_c(l|\lambda)$. This probability can be estimated from the distribution of objects and rooms in the training dataset.

Next, we use $p_c(l|\lambda)$ to estimate the probability of different locations $l$ accomplishing the goal. Since $M$ predicted by `ARGoal` includes the objects potentially involved in the subgoal and their corresponding probabilities, the probability of achieving the goal in different locations within the environment can be calculated using Bayes' theorem. Therefore, the probability of accomplishing the task by moving to location $l$ is:

$$p(l|M) = \sum_{\lambda \in M} p_c(l|\lambda) \cdot p_{\theta_1}(\lambda). \tag{14}$$

Based on $p(l|M)$, we can construct a heuristic environment search sequence. when the LLM outputs a repeated or unreasonable $l$, the next location will be sampled sequentially from the unvisited locations in the above list.

## 4 Experiments

### 4.1 Datasets and Evaluation Metrics

We conducted experiments on the HandMeThat Dataset Version 2 [39]. This version expands the data size by a factor of 40 compared to its initial release, focusing on a narrower set of goals while generating larger sampling space of world states. The version 2 dataset comprises a total of 116,146 episodes, with each goal template represented by approximately 5,000 episodes. The distribution of episodes across different hardness levels are 4,008, 7,8488, 2,8343, and 5,307, respectively. The target object at simpler levels (level 1) can be directly grounded from the instructions. However, as the difficulty level escalates, the agent must engage in pragmatic reasoning and goal inference to locate the target object. The abundance of data for higher difficulty levels allows for a more comprehensive evaluation of model performance in goal inference and pragmatic reasoning. Three evaluation metrics are considered: 1) average score (AS): it requires the agent with high success rate but less action cost; 2) success rate (SR): the agent's ability to help human achieves the goal within limited steps (i.e., 40 steps); 3) action cost (AC): average number of actions in successful episodes. We repeat each experiment three times and calculate the standard deviation.

### 4.2 Implementation Details

We employ `Llama-2-7b` [35] as backbone models and conduct experiments using the `llama-recipes`[3] repository. length of 4096 across all scales. We fine-tuned the model for one epoch using LoRA [17], with a learning rate of $1 \times 10^{-4}$ and a sequence length of 4096. We train the model using eight A100 GPUs with 80GB of memory each and a batch size of four per GPU over a period of about eight hours. As for the GPT-4 [1] baseline, we access the `gpt-4-turbo-preview` model from the OpenAI API and use the parameter of temperature 0.7. We employed a mixture of three types of data generated from expert episodes to fine-tune our model, which includes data for action prediction planning, data

[3] https://github.com/meta-llama/llama-recipes/tree/main

| Method | Metrics | Partially Observable | | | |
|---|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 | Level 4 |
| Human | AS↑ | - | - | - | - |
| | SR↑ | 100% | 80% | 40% | 30% |
| | AC↓ | 7.8 | 7.5 | 8.4 | 8.0 |
| Random | AS↑ | -40.0 | -39.5 | -40.0 | -40.0 |
| | SR↑ | 0.0% | 0.4% | 0.0% | 0.0% |
| | AC↓ | N/A | 16.0 | N/A | N/A |
| Seq2Seq | AS↑ | 1.3 | 2.4 | -17.7 | -23.7 |
| | SR↑ | 30.4% | 31.2% | 16.4% | 12.0% |
| | AC↓ | 4.0 | **4.0** | 4.1 | 4.0 |
| Seq2Seq +goal | AS↑ | 0.6 | 3.0 | -9.6 | -9.6 |
| | SR↑ | 29.8% | 31.6% | 22.4% | 22.4% |
| | AC↓ | **4.0** | **4.0** | 4.1 | 4.1 |
| Seq2Seq +subgoal | AS↑ | 0.6 | 14.4 | -1.9 | -1.4 |
| | SR↑ | 29.8% | 40.0% | 28.0% | 28.4% |
| | AC↓ | **4.0** | **4.0** | **4.0** | **4.0** |
| DRRN | AS↑ | -40.0 | -40.0 | -40.0 | -40.0 |
| | SR↑ | 0.0% | 0.0% | 0.0% | 0.0% |
| | AC↓ | N/A | N/A | N/A | N/A |
| GPT-4 +KnowAgent | AS↑ | 57.24 | 36.28 | -7.16 | 4.08 |
| | SR↑ | 64.0% | 44.0% | 0.0% | 12.0% |
| | AC↓ | 5.31 | 5.09 | N/A | 6.67 |
| **IEKP (Ours)** | AS↑ | **91.97**±0.58 | **83.36**±0.16 | **68.95**±0.80 | **72.11**±2.99 |
| | SR↑ | **98.89%**±0.45% | **93.48%**±0.00% | **85.69%**±0.01% | **87.82%**±0.03% |
| | AC↓ | 6.55±0.19 | 7.97±0.10 | 12.18±0.49 | 12.18±0.49 |

**Table 1.** Experiment results in partially observable setting. Each model is evaluated on 4 hardness levels with 3 metrics: average score (AS, higher is better), success rate (SR, higher is better), and action cost (AC, lower is better).

for achieving human goal recognition, and data for predicting task categories. In the paper, $\theta_{0,1,2,3,4}$ refer to the same LLM implemented with different prompts. The training data for action type prediction, target object prediction, and action template generation are summarized from the HandmeThat training set.

### 4.3 Quantitative Experiments

The evaluation of methods on HandMeThat [39] encompasses three distinct sets: **1) Hand-coded models**. The set of hand-coded models comprises a random agent ("Random") and an agent based on heuristic rules ("Heuristic"). Specifically, the random agent selects a valid action at random at each time step. In contrast, the heuristic agent, which operates exclusively in a fully-observable environment, has access to the states of all objects and the logic formula underlying the utterance. The central heuristic of this model posits that humans are inclined to seek objects belonging to the same categories as those they have previously interacted with. **2) Neural network models.** We evaluate two neural network baselines, which have been trained utilizing both offline and online reinforcement learning algorithms. The first model ("Seq2Seq") is developed based on the sequence-to-sequence architecture [34] and trained through imitation learning, utilizing expert demonstrations generated by FF heuristic [16]. The second model, referred to as "DRRN", is an agent designed for choice-based text games [15]. **3) GPT-4 [1] based model [44].** We equiped GPT-4 with the prompt for ALFWorld in KnowAgent [44], a state-of-the-art planning method for LLM-Based Agents.

| Method | Metrics | Fully Observable | | | |
|---|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 | Level 4 |
| Human | AS↑ | - | - | - | - |
| | SR↑ | 100% | 80% | 40% | 30% |
| | AC↓ | 4.1 | 4.2 | 5.2 | 4.5 |
| Heuristic | AS↑ | 95.8 | 59.5 | 34.5 | 24.4 |
| | SR↑ | 100.0% | 64.0% | 39.2% | 29.2% |
| | AC↓ | 4.2 | 4.1 | 4.1 | 4.1 |
| Random | AS↑ | -40.0 | -39.8 | -40.0 | -40.0 |
| | SR↑ | 0.0% | 0.1% | 0.0% | 0.0% |
| | AC↓ | N/A | 30.0 | N/A | N/A |
| Seq2Seq | AS↑ | 1.3 | -0.8 | -22.6 | -19.9 |
| | SR↑ | 30.4% | 28.8% | 12.8% | 14.8% |
| | AC↓ | **4.0** | **4.0** | **4.0** | **4.0** |
| Seq2Seq +goal | AS↑ | -4.7 | 7.3 | -6.9 | -6.9 |
| | SR↑ | 26.0% | 34.8% | 24.4% | 24.4% |
| | AC↓ | **4.0** | 4.1 | 4.1 | 4.1 |
| Seq2Seq +subgoal | AS↑ | 53.8 | 51.3 | 1.3 | 3.0 |
| | SR↑ | 69.1% | 67.2% | 30.4% | 31.6% |
| | AC↓ | 4.2 | 4.1 | 4.1 | 4.1 |
| DRRN | AS↑ | -40.0 | -40.0 | -40.0 | -40.0 |
| | SR↑ | 0.0% | 0.0% | 0.0% | 0.0% |
| | AC↓ | N/A | N/A | N/A | N/A |
| **IEKP (Ours)** | AS↑ | **94.49**±0.34 | **86.09**±0.24 | **73.00**±1.17 | **72.57**±0.61 |
| | SR↑ | **99.37%**±0.00% | **93.67%**±0.00% | **86.39%**±0.01% | **86.67%**±0.00% |
| | AC↓ | 4.65±0.05 | 5.31±0.05 | 9.04±0.08 | 9.98±0.10 |

**Table 2.** Experiment results in fully observable setting. Each model is evaluated on 4 hardness levels with 3 metrics: average score (AS, higher is better), success rate (SR, higher is better), and action cost (AC, lower is better).

The quantitative results in partially observable and fully observable settings are shown in Table 1 and Table 2, respectively. The Random and DRRN agents struggle significantly to achieve the subgoals specified by humans in both settings, primarily due to the extensive action space and the sparsity of reward feedback. The study presented in [39] restricts human participants to manipulating a single object and records 25 episodes of performance per level of difficulty, denoted as "Human" performance. The decline in performance with rising hardness levels confirms that progressively ambiguous instructions make it challenging for human to infer the speaker's intentions with only one attempt. In fully observable setting, the experiment results of the Heuristic agent are comparable to human performance. Through behavior cloning, the Seq2Seq agent can explore a wealth of expert training data. Incorporating additional oracle information, i.e., goal or subgoal, further enhances the Seq2Seq model's performance. For the GPT-4 agent, we collected data on the performance across randomly sampled 25 episodes for each level of difficulty. It demonstrates exceptional performance at elementary levels due to its robust zero-shot reasoning and planning capabilities. However, the performance of the GPT-4 agent significantly decreases due to hallucinations (e.g., wrong association and non-existent object) when subgoals are under-specified.

Compared to previous methods, IEKP demonstrates a significant accuracy advantage in both fully and partially observable settings. In the fully observable setting, at Level 1 difficulty, our method's accuracy is comparable to that of the Heuristic approach, with both achieving nearly 100% SR. However, as the

difficulty increases and the information within the instructions becomes more ambiguous, our method exhibits a clear advantage over others. At Level 4 difficulty, our method achieves an accuracy of 86.67% in SR, surpassing previous methods. Furthermore, the accuracy of our method declines minimally as the difficulty escalates from Level 1 to Level 4, indicating a marked improvement in the success rate of intent comprehension to previous methods. Similarly, in the partially observable setting, our method shows significant improvement and outperforms gpt-4 based models, thereby validating the efficacy of our proposed approach.

**Table 3.** Verifying the effectiveness of each component in our proposed IEKP mechanism.

| FDPrune | ARGoal | K-Search | Level 1 | | | Level 2 | | | Level 3 | | | Level 4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | AS↑ | SR↑ | AC↓ | AS↑ | SR↑ | AC↓ | AS↑ | SR↑ | AC↓ | AS↑ | SR↑ | AC↓ |
| | | | 74.30 | 84.33 | **4.47** | 56.52 | 71.33 | **4.69** | 11.09 | 37.67 | **4.37** | 16.24 | 41.67 | **5.02** |
| ✓ | | | 88.52 | 95.5 | 5.42 | 80.79 | 90.83 | 7.02 | 61.81 | 78.67 | 10.54 | 58.84 | 76.17 | 10.21 |
| ✓ | ✓ | | 90.07 | 96.67 | 5.44 | 79.76 | 90.17 | 7.17 | 63.32 | 80.17 | 11.09 | 61.09 | 77.83 | 10.10 |
| ✓ | | ✓ | 92.15 | 98.83 | 6.27 | 82.70 | 92.67 | 7.51 | 68.09 | 84.50 | 11.88 | 66.72 | 83.33 | 11.77 |
| ✓ | ✓ | ✓ | **92.62** | **99.17** | 6.26 | **84.67** | **94.33** | 7.80 | **74.01** | **89.67** | 12.74 | **69.68** | **86.00** | 12.30 |

### 4.4 Ablation Study

To assess the varying designs within our framework, we conducted ablation studies for each level of difficulty. Tab. 3 presents the ablation results of our proposed modules. The 1-st row indicates the baseline method based on LLM, which is fine-tuned on the training expert data through imitation learning.

**About `FDPrune`.** As illustrated in Tab. 3, compared to #1, the model with `FDPrune` exhibits an increase in both AS and SR from level 1 to level 4 difficulty. The model with `FDPrune` exhibited a 14.22 increase in AS at level 1, with a particularly notable enhancement at level 4, where AS surged by 42.60. This indicates that `FDPrune` significantly enhances the robustness of decision-making, especially when instructions from human are ambiguous, necessitating multiple backtracking explorations during the planning process, by leveraging finite state machine pruning.

**About `ARGoal`.** Compared to line #2, we input the $M$ predicted by `ARGoal` into the LLM through CoT [40] to assist in inferring subsequent actions at #3. When utilizing `ARGoal`, the model achieved performance improvements at level 1 (AS 1.55 ↑), level 2 (AS 1.51 ↑), and level 4 (AS 2.25 ↑). This indicates that goal recognition information provided by `ARGoal` is effective for executing ambiguous instructions. When used in conjunction with `K-search`, `ARGoal` significantly enhances the performance of the model, as demonstrated below.

**About `K-Search`.** When the `K-search` employs the goal representations $M$ predicted by `ARGoal`, the model's accuracy experiences a notable enhancement. Compared to the scenario where only `ARGoal` is utilized, the incorporation of `K-search` significantly elevates the precision at level 4, with 5.63

AS improvement. This underscores the substantial improvement in the agent's goal-searching capabilities facilitated by co-occurrence relationships, particularly when the task instructions are ambiguous. The interpretable goal-search strategy based on `K-search`, as opposed to the implicit goal provision by CoT [40] influencing decision-making, exhibits superior robustness.

**About Association Process of `ARGoal`.** `ARGoal` initially estimates the goal with LLM, designated as $M_1$ in §3.3. Subsequently, `ARGoal` employs substitution relationships to conjecture additional objects, thereby enriching the final goal recognition outcome $M$. Consequently, in Experiment #4, we eliminated the associative process within `ARGoal`, relying solely on $M_1$ to conduct the `K-search`, in order to validate the efficacy of the association process. Compared to line #5, the success rate at level 4 decreased by 2.67%. This indicates that association can significantly enhance search efficiency and improve the expressive power of $M$.

## 5 Conclusion and Discussion

In this paper, we explore the task of human-AI collaboration in domestic settings, particularly focusing on ambiguous commands, based on the handmeThat [39]. We introduce the IEKP model, which has achieved state-of-the-art accuracy on benchmarks, surpassing even human performance in complex scenarios, which has three primary phases. First, `ARGoal` predicts human intent representations based on objects and enhances the selection by incorporating associative capabilities to include alternative objects. Second, `FDPrune` predicts the type of task and constructs a finite state machine to prune illogical actions during the decision-making process. Third, `K-search` enhances the agent's efficiency in searching for targets within the environment by leveraging object co-occurrence relationships. Our ablation studies confirm the effectiveness of each componets of IEKP. IEKP significantly improves the understanding of unspecific utterance within a text-based action space, offering substantial prospects for applications such as domestic service robots and elderly care services.

*Future work.* Given that the methodology discussed in this article revolves around abstracting the action space for textual environments, there exists a discernible gap with real, embodied scenarios. Consequently, future work could focus on the study of human comprehension and execution of ambiguous instructions under embodied visual perception. Additionally, during this collaborative process, communication between humans and machines is unidirectional; when the robot is confused about the instructions, humans are unable to provide further guidance, nor can the robot pose questions to the humans. Therefore, exploring this interaction process also merits attention in future endeavors.

*Acknowledgments* This research is supported in part by National Key R&D Program of China (2022ZD0115502), National Natural Science Foundation of China (NO.62461160308, U23B2010), "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2024C01161).

## References


1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3674–3683 (2018)
3. Baker, C.L., Tenenbaum, J.B., Saxe, R.: Goal inference as inverse planning. CogSci (2007)
4. Bao, J., Duan, N., Zhou, M., Zhao, T.: Knowledge-based question answering as machine translation. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 967–976 (2014)
5. Chen, J., Gao, C., Meng, E., Zhang, Q., Liu, S.: Reinforced structured state-evolution for vision-language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15450–15459 (2022)
6. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. Journal of Machine Learning Research **25**(70), 1–53 (2024)
7. Dennett, D.C.: The intentional stance. MIT press (1987)
8. Ding, Z., Ding, Z.h., Hui, T., Huang, J., Wei, X., Wei, X., Liu, S.: Ppmn: Pixel-phrase matching network for one-stage panoptic narrative grounding. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 5537–5546 (2022)
9. Ding, Z., Hui, T., Huang, J., Wei, X., Han, J., Liu, S.: Language-bridged spatial-temporal interaction for referring video object segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4964–4973 (2022)
10. Gallagher, H.L., Frith, C.D.: Functional imaging of 'theory of mind'. Trends in cognitive sciences **7**(2), 77–83 (2003)
11. Gao, C., Chen, J., Liu, S., Wang, L., Zhang, Q., Wu, Q.: Room-and-object aware knowledge reasoning for remote embodied referring expression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3064–3073 (2021)
12. Gao, C., Liu, S., Chen, J., Wang, L., Wu, Q., Li, B., Tian, Q.: Room-object entity prompting and reasoning for embodied referring expression. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(2), 994–1010 (2023)
13. Geffner, H., Bonet, B.: A concise introduction to models and methods for automated planning. Morgan & Claypool Publishers (2013)
14. Gergely, G., Nádasdy, Z., Csibra, G., Bíró, S.: Taking the intentional stance at 12 months of age. Cognition **56**(2), 165–193 (1995)
15. He, J., Chen, J., He, X., Gao, J., Li, L., Deng, L., Ostendorf, M.: Deep reinforcement learning with an action space defined by natural language (2016)
16. Hoffmann, J., Nebel, B.: The ff planning system: Fast plan generation through heuristic search. Journal of Artificial Intelligence Research **14**, 253–302 (2001)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)

18. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232 (2023)
19. Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Gordon, D., Zhu, Y., Gupta, A., Farhadi, A.: AI2-THOR: An Interactive 3D Environment for Visual AI. arXiv:1712.05474 (2017)
20. Kong, X., Chen, J., Wang, W., Su, H., Hu, X., Yang, Y., Liu, S.: Controllable navigation instruction generation with chain of thought prompting. In: European Conference on Computer Vision. pp. 37–54. Springer (2024)
21. Levesque, R.J.R.: Social Reasoning, pp. 2808–2808. Springer New York (2011)
22. Li, X., Zhao, R., Chia, Y.K., Ding, B., Bing, L., Joty, S., Poria, S.: Chain of knowledge: A framework for grounding large language models with structured knowledge bases. arXiv preprint arXiv:2305.13269 (2023)
23. Meneguzzi, F., Pereira, R.: A survey on goal recognition as planning. In: IJCAI (2021)
24. Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., et al.: Augmented language models: a survey. arXiv preprint arXiv:2302.07842 (2023)
25. Puig, X., Ra, K., Boben, M., Li, J., Wang, T., Fidler, S., Torralba, A.: Virtualhome: Simulating household activities via programs. In: CVPR (2018)
26. Puig, X., Shu, T., Li, S., Wang, Z., Liao, Y.H., Tenenbaum, J.B., Fidler, S., Torralba, A.: Watch-and-help: A challenge for social perception and human-{ai} collaboration. In: ICLR (2021)
27. Qiao, S., Gui, H., Chen, H., Zhang, N.: Making language models better tool learners with execution feedback. arXiv preprint arXiv:2305.13068 (2023)
28. Ramírez, M., Geffner, H.: Plan recognition as planning. In: IJCAI (2009)
29. Ramírez, M., Geffner, H.: Probabilistic plan recognition using off-the-shelf classical planners. In: AAAI (2010)
30. Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., Fox, D.: Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10740–10749 (2020)
31. Shridhar, M., Yuan, X., Cote, M.A., Bisk, Y., Trischler, A., Hausknecht, M.: Alfworld: Aligning text and embodied environments for interactive learning. In: International Conference on Learning Representations (2020)
32. Stenmark, M., Malec, J.: Knowledge-based instruction of manipulation tasks for industrial robotics. Robotics and Computer-Integrated Manufacturing **33**, 56–67 (2015)
33. Suhr, A., Yan, C., Schluger, J., Yu, S., Khader, H., Mouallem, M., Zhang, I., Artzi, Y.: Executing instructions in situated collaborative interactions. In: EMNLP-IJCNLP (2019)
34. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. Advances in neural information processing systems **27** (2014)
35. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
36. Trivedi, H., Balasubramanian, N., Khot, T., Sabharwal, A.: Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. arXiv preprint arXiv:2212.10509 (2022)

37. Urbanek, J., Fan, A., Karamcheti, S., Jain, S., Humeau, S., Dinan, E., Rocktäschel, T., Kiela, D., Szlam, A., Weston, J.: Learning to speak and act in a fantasy text adventure game. In: EMNLP-IJCNLP (2019)
38. Vu, T., Iyyer, M., Wang, X., Constant, N., Wei, J., Wei, J., Tar, C., Sung, Y.H., Zhou, D., Le, Q., et al.: Freshllms: Refreshing large language models with search engine augmentation. arXiv preprint arXiv:2310.03214 (2023)
39. Wan, Y., Mao, J., Tenenbaum, J.B.: Handmethat: Human-robot communication in physical and social environments. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022)
40. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
41. Zhao, R., Chen, H., Wang, W., Jiao, F., Do, X.L., Qin, C., Ding, B., Guo, X., Li, M., Li, X., et al.: Retrieving multimodal information for augmented generation: A survey. arXiv preprint arXiv:2303.10868 (2023)
42. Zhao, Y., Chen, J., Gao, C., Wang, W., Yang, L., Ren, H., Xia, H., Liu, S.: Target-driven structured transformer planner for vision-language navigation. In: Proceedings of the 30th ACM international conference on multimedia. pp. 4194–4203 (2022)
43. Zhi-Xuan, T., Mann, J., Silver, T., Tenenbaum, J., Mansinghka, V.: Online bayesian goal inference for boundedly rational planning agents. In: NeurIPS (2020)
44. Zhu, Y., Qiao, S., Ou, Y., Deng, S., Zhang, N., Lyu, S., Shen, Y., Liang, L., Gu, J., Chen, H.: Knowagent: Knowledge-augmented planning for llm-based agents. arXiv preprint arXiv:2403.03101 (2024)