Situated Instruction Following Under Ambiguous Human Intent

Anonymous Author(s)
Affiliation
Address
email

Abstract

Language is never spoken in a vacuum. It is expressed and comprehended within the holistic backdrop of the speaker's history, actions, and environment. Since humans are used to communicating efficiently with situated language, the practicality of robotic assistants hinge on their ability to understand and act upon implicit and situated instructions. In traditional instruction following paradigms, the agent acts alone in an empty house, leading to language use that is both simplified and artificially "complete." In contrast, we propose *situated instruction following* (SIF), which embraces the inherent underspecification and ambiguity of real-world communication with the physical presence of a human speaker. The meaning of situated instructions naturally unfold through the past actions and the expected future behaviors of the human involved. Specifically, within our settings we have instructions that (1) are ambiguously specified, (2) have temporally evolving intent, (3) can be interpreted more precisely with the agent's dynamic actions. Our experiments indicate that state-of-the-art Embodied Instruction Following (EIF) models lack holistic understanding of situated human intention.

1 Introduction

Humans communicate efficiently by providing only the necessary information, relying on shared context like history, actions, and environment. For example, the request "Can you bring me a cup?" varies based on context—if said near a kitchen sink with gloves, it likely refers to a dirty cup, while near a bathroom sink, it suggests a clean one. Although clarification is possible, humans often interpret such requests accurately using contextual cues, showing our ability to derive nuanced, situation-specific meanings from ambiguous language.

As robotic agents increasingly become integral to our daily lives, their effectiveness and utility critically depend on their ability to comprehend and respond to situated language— natural language spoken by humans. Without this capability, agents may prove more of a hindrance than a help, forcing users to perform tasks themselves rather than entrusting them to an assistant. As discussed in the field of agent alignment [Leike et al., 2018], it is often difficult for users to precisely define or articulate ideal task specifications. Consequently, an agent that demands detailed explanations might render manual task execution by humans more attractive.

Current instruction-following tasks prioritize accurate low-level instruction interpretation [Anderson et al., 2018, Gu et al., 2022, Padmakumar et al., 2021, Shridhar et al., 2020] or use commonsense to achieve underspecified goals like object navigation [Chaplot et al., 2020, Das et al., 2018]. In contrast, our work SIF aims to generalize *Embodied* Instruction Following to *Situated* Instruction Following, with instructions closer to the language naturally spoken by humans. Specifically, we focus on three dimensions of situated reasoning:

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

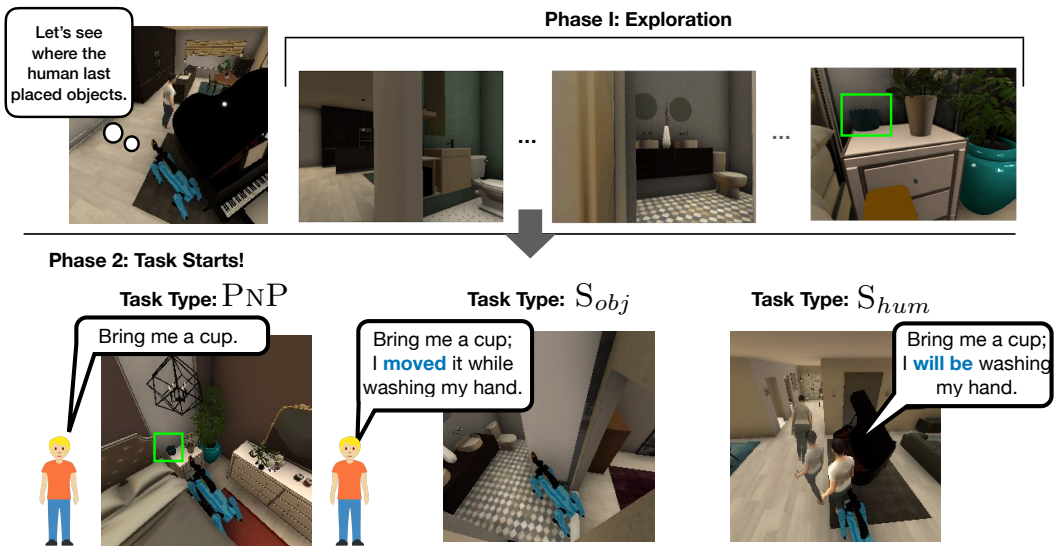

Figure 1: **Overview.** The tasks in SIF consist of two phases: an exploration phase (phase 1) and a task phase (phase 2). PNP represents a conventional static Pick-and-Place task used for comparison, wherein the environment remains unchanged after the exploration phase. S_{hum} and S_{obj} introduce two novel types of situated instruction following tasks. In these tasks, the *objects* and *human* subjects move during the task phase. Nuanced communication regarding these movements is provided, necessitating reasoning about ambiguous and temporally evolving human intent.

1. **Ambiguity:** As in the cup example above, there is ambiguity in the instruction given by the speaker.

2. **Temporal:** A speaker's actions change how their instruction should be interpreted (e.g., clarifying an underspecified reference).

3. **Dynamic:** When the environment changes, the agent needs to decide what actions will reduce their uncertainty (e.g., following the human).

We implement our tasks in Habitat 3.0 [Puig et al., 2023], which includes simulated human agents. To ensure fair comparison with prior work, we include both static (prior work) and dynamic (this work) tasks (Fig. 1). The static task follows the classic pick-and-place paradigm where the agent is instructed to `Put [Obj] in/on [Recep]`. We simplify the setup by allowing the agent to explore, minimizing the role of mapping in our reasoning benchmark.

Our benchmark focuses on dynamic tasks where the agent must combine instruction understanding with human movement. The dynamic tasks include S_{obj}(object moved by human) and S_{hum}(human is the receptacle). In these, the agent receives goal instructions (e.g., "Bring me a mug" for S_{hum}or "Put the mug in the bathroom" for S_{obj}) along with relocation hints. In S_{hum}, the human moves as the task begins, signaling intent through both words and movement. The agent must efficiently follow instructions, retrieve the object, and place it in the correct location (e.g., with the moving human in S_{hum}tasks).

We specifically target evaluation of state-of-the-art Embodied Instruction Following (EIF) baselines. We implement two such systems inspired by papers on related tasks. The first baseline, which we refer to as REASONER, is a closed-loop system incorporating a semantic map, a prompt generator, and a Large Language Model (LLM) planner. For the prompt generator, we integrated components from Voyager[Wang et al., 2023], LLMPlanner[Song et al., 2023], and ReAct[Yao et al., 2022], tailoring them to suit our dataset's specific requirements. The second baseline, PROMPTER[Inoue and Ohashi, 2022], was very successful at executing ALFRED [Shridhar et al., 2020] tasks despite being open-loop. We see the desired result that our static scenarios match those from existing EIF datasets [Inoue and Ohashi, 2022, Song et al., 2023], and these LLM based approaches perform very well in tasks requiring common sense. However, their performance significantly declines when faced with situations that require reasoning about the human's behavior.

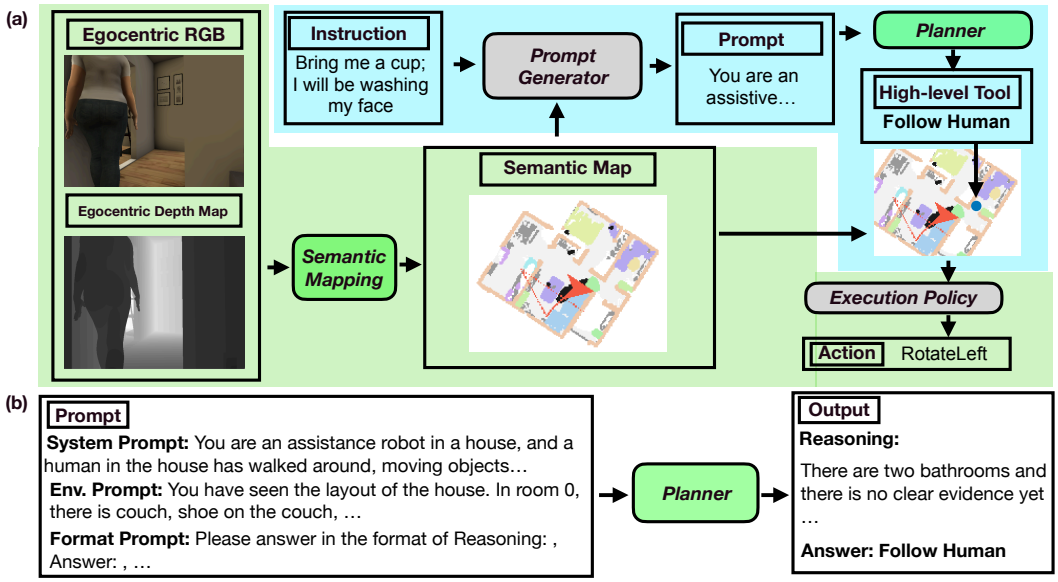

Figure 2: REASONER: (a) The semantic mapper is updated at every timestep, whereas the prompt generator and planner are activated either upon completion of the last high-level action or when a new decision is required. (b) The prompt consists of system prompt, environment prompt, format prompt.

2 Task

Our tasks (1) are structured into two distinct phases: (1) the exploration phase and (2) the task phase. During the exploration phase, the agent is allotted N steps to navigate around a static house environment where object assets are positioned. The value of N is determined to ensure the agent has sufficient steps to thoroughly scan the environment; specifically, N = 1.5 x (the number of steps required to achieve a complete map using frontier-based exploration techniques). Following the exploration phase, some objects are repositioned without the agent's knowledge. As the task phase commences, the agent receives an instruction (e.g., "Bring me a cup," "Put the cup in the sink"), accompanied by either direct or ambiguous information regarding which objects have been moved (e.g., "I took a cup with me. I'll be getting ready for bed"). If the task involves delivering an object to a human, the human walks into the agent's field of view as the task begins, simultaneously providing hints about their intended location ("I will be in the bathroom washing my face"). These elements, along with other strategic design decisions, ensure that the exploration phase effectively contextualizes the language directives, rendering tasks sufficiently solvable.

3 Baselines

Many recent state-of-the-art EIF agents are modular models with an LLM planner, connected to learned/engineered episodic memory, perception, and execution tools. We present a baseline within this high-performing family — REASONER, a closed-loop baseline that adapts FILM[Min et al., 2021] and the prompts of llm-planner[Song et al., 2023], and ReAct [Yao et al., 2022], and prompter [Inoue and Ohashi, 2022], an open-loop SOTA agent built for ALFRED [Shridhar et al., 2020].

Semantic Mapper. The semantic mapper creates a global representation for visual observation. As in previous work[Chaplot et al., 2020, Min et al., 2021], we process egocentric RGB and depth into an allocentric top-down map of obstacles and semantic categories using Detic[Zhou et al., 2022]. The semantic categories of interest are `[ObjectCat]`, `[Recep]`, and "human." In contrast to previous works[Chaplot et al., 2020, Min et al., 2021], the most recent human and object positions are refreshed post new observations and pick/place actions, ensuring a dynamic and accurate representation of the environment.

Text representation generator. The semantic map and other contexts are converted into prompts. It is a concatenation of three components: the system prompt, environment prompt, and the format prompt:

Table 1: **SPL** performance of REASONER across splits. In each sectioned-row, the top row assumes oracle perception (semantic segmentation and manipulation); the bottom row assumes learned semantic segmentation and heuristic manipulation. To minimize the burden on API costs and time, we have limited LLM API calls for plan generation to 15 times.

Model		Val Seen			Val Unseen			Test Seen			Test Unseen		
Planning	Perception	PNP	S_{obj}	S_{hum}	PNP	S_{obj}	S_{hum}	PNP	S_{obj}	S_{hum}	PNP	S_{obj}	S_{hum}
Oracle	Oracle	98	100	95	100	100	100	98	93	98	95	100	98
	Learned	46	46	59	41	30	54	52	30	69	44	47	46
REASONER	Oracle	82	61	23	78	49	39	73	58	29	81	49	34
	Learned	21	8	12	24	11	12	29	2	15	18	14	15

- **System:** The system prompt outlines the agent's role and and encourages it to account for uncertainty. It is presented as "You are an assistive robot in a house, aiding a human. Your observations may be incomplete or wrong."
- **Environment:** The environment prompt is a conversion of the episodic memory into text format, and contains information of the agent's current state and previously completed/failed actions. It is given in the following sequence: (1) observation of P_e during exploration phase, based on the semantic map, (2) C, regarding object/ human movements, (3) the goal instruction I, (4) the high-level action executed by agents at timesteps and their observed consequences (success/fail), (5) the agent's latest observation, based on the semantic map.
- **Format:** The format prompt explains action affordance and a format for chain of thought [Wei et al., 2022]. It also explains the desired effect of actions (e.g. "If you want to keep searching for object(s) or human that might exist (but you have not detected) in the current room, choose 'Explore Room X' (Table 2).")

Execution Tools Upon receiving the prompt, the planner is prompted to choose a high-level action (Tab. 2); then corresponding execution tools are called. A complete list of tools are listed in Table 2. When the execution is done, the tool sends this message, and the prompt generator creates a new prompt and the planner calls a new tool.

4 Results

Results from our experiments are presented in Table 1. This table notably shows the following facts about our dataset and baselines. First, the gap of model performance across PNP versus S_{hum}, S_{obj} shows that PNP can be solved with commonsense and mechanistic combination, and the rest two tasks cannot. The reasoning challenges of S_{obj} and S_{hum} are backed by the performance of REASONER with oracle perception/manipulation; it shows a stark contrast in PNP tasks ($\sim 80\%$) and S_{obj}, S_{hum} tasks ($\sim 45\%$).

Table 3 examines model performance on clear versus ambiguous tasks. Ambiguity in S_{obj} tasks emerges when multiple potential locations exist for an object, as indicated by communicative cues. For example, the statement "I am washing my face" becomes ambiguous when multiple bathrooms are available. Similarly, ambiguous S_{hum} tasks occur when the human could be in several different locations. In S_{hum} tasks, REASONER underperforms in clear tasks due to a tendency to conservatively judge that there is insufficient evidence of the human's destination, even when only one plausible location exists. REASONER attempts some calibration but generally leans towards following the human. Qualitative analysis reveals that in ambiguous tasks, REASONER often disengages prematurely, assuming it has accumulated enough evidence.

5 Conclusion

We present Situated Instruction Following (SIF), a new dataset to evaluate situated and holistic understanding of language instructions. Our dataset reflects aspects of real-world instruction following: (1) ambiguous task specification, (2) evolving intent over time, and (3) dynamic interpretation influenced by agent action. We show that current state-of-the-art models struggle with this level of understanding, further highlighting the complexity and uniqueness of our dataset.

References

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018.

Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33, 2020.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2018.

Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Eric Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. *arXiv preprint arXiv:2203.12667*, 2022.

Yuki Inoue and Hiroki Ohashi. Prompter: Utilizing large language model prompting for a data efficient embodied instruction following. *arXiv preprint arXiv:2211.03267*, 2022.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

So Yeon Min, Devendra Singh Chaplot, Pradeep Ravikumar, Yonatan Bisk, and Ruslan Salakhutdinov. Film: Following instructions in language with modular methods. *arXiv preprint arXiv:2110.07342*, 2021.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat, 2021. URL `https://arxiv.org/abs/2110.00534`.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*, 2023.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009, 2023.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022.

A Execution Tools

Execution tools for REASONER/PROMPTER and their working details/affordance are in Table 2.

Execution Tool	Description & Affordance
Navigation	
Go to Room X	FMM Planner navigates to a random point in Room X.
Explore Room X	FMM Planner navigates to a random point in Room X; then, agent turns 15 times to the right to look around.
Follow Human	The last observed position of the human is given as the goal, to the human-following wrapper (more explanation is Sec. **??**) on top of FMM Planner.
Manipulation	
Grab Obj	The closest object within 2 meters of the grasper is grabbed, and agent's grasper is closed.
Put Obj	Grasped object is put on the closest receptacle within 2 meters of the grasper is grabbed, and agent's grasper is opened.
Give Obj to Human	The agent goes within 1 meter of the human and gives grasped object to human, if human is visible from current view.

Table 2: Execution tools for REASONER/PROMPTER and their working details/affordance.

B Ablations Across Ambiguous/ Clear Tasks

Table 3: **Ambiguous vs Clear tasks**. SPL and SR of REASONER and PROMPTER with G.T./learned vision and manipulation on Val seen & unseen combined.

Model	**Metric**	**G.T. Vis. & Man.**				**Learned Vis. & Man.**			
		$\mathbf{S}_{obj}$		$\mathbf{S}_{hum}$		$\mathbf{S}_{obj}$		$\mathbf{S}_{hum}$	
		Clear	Amb.	Clear	Amb.	Clear	Amb.	Clear	Amb.
REASONER	SPL	62	52	13	42	9	11	3	17
	SR	76	71	14	67	15	14	6	26
PROMPTER	SPL	38	29	3	42	11	8	0	17
	SR	54	36	4	66	18	10	0	27