# ACKnowledge: A Computational Framework for Human Compatible Affordance-based Interaction Planning in Real-world Contexts

Ziqi Pan
The Hong Kong University of Science and Technology
Hong Kong, Hong Kong
zpanar@connect.ust.hk

Xiucheng Zhang
School of Artificial Intelligence
Sun Yat-sen University
Zhuhai, China
zhangxch58@mail2.sysu.edu.cn

Zisu Li
IIP (Computational Media and Arts)
The Hong Kong University of Science and Technology
Hong Kong SAR, Hong Kong, China
zlihe@connect.ust.hk

Zhenhui Peng
School of Artificial Intelligence
Sun Yat-sen University
Zhuhai, Guangdong Province, China
pengzhh29@mail.sysu.edu.cn

Mingming Fan
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
mingmingfan@ust.hk

Xiaojuan Ma
The Hong Kong University of Science and Technology
Hong Kong, Hong Kong
mxj@cse.ust.hk

## Abstract

Intelligent agents coexisting with humans often need to interact with human-shared objects in environments. Thus, agents should plan their interactions based on objects' affordances and the current situation to achieve acceptable outcomes. How to support intelligent agents' planning of affordance-based interactions compatible with human perception and values in real-world contexts remains under-explored. We conducted a formative study identifying the physical, intrapersonal, and interpersonal contexts that count to household human-agent interaction. We then proposed *ACKnowledge*, a computational framework integrating a dynamic knowledge graph, a large language model, and a vision language model for affordance-based interaction planning in dynamic human environments. In evaluations, *ACKnowledge* generated acceptable planning results with an understandable process. In real-world simulation tasks, *ACKnowledge* achieved a high execution success rate and overall acceptability, significantly enhancing usage-rights respectfulness and social appropriateness over baselines. The case study's feedback demonstrated *ACKnowledge*'s negotiation and personalization capabilities toward an understandable planning process.

## CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**; *Ambient intelligence.*

## Keywords

Affordance-based Interaction Planning, Real-world Context, Human Compatible


*CHI '25, Yokohama, Japan*

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713791

**ACM Reference Format:**
Ziqi Pan, Xiucheng Zhang, Zisu Li, Zhenhui Peng, Mingming Fan, and Xiaojuan Ma. 2025. ACKnowledge: A Computational Framework for Human Compatible Affordance-based Interaction Planning in Real-world Contexts. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3706598.3713791

## 1 Introduction

Intelligent agents are increasingly integrated into various aspects of human life, including classrooms [33, 45], hospitals [5, 124], workplaces [39], and households [46]. As assistants, collaborators, or companions [119], these agents are expected to interact with physical entities shared or owned by humans. In shared environments with different contexts, the agents should follow how humans normally perceive and reason about [1, 34, 97] the interaction possibilities an environment offers — i.e., the affordances of objects [41] — and act [74] accordingly to be compatible with humans. Failing to comply with human practices physically, psychologically, and socially [28] may damage objects, disrupt environments, and consequently downgrade humans' experience of interacting with the agents [17, 28, 48, 62].

While human compatibility through explainability [109] and personalizability [64] is well recognized in Human-Robot Interaction (HRI) literature, our work aims to extend these existing concepts and offer new insights into practically achieving it when agents plan their high-level interactions with objects in dynamic human environments with complex, implicit constraints. First, we emphasize that **an agent should understand and model how humans perceive and reason about the affordances of objects in dynamic environments** to be in tune with human activities. Second, we highlight that **reconciliation with human practices should be enhanced in every intended move of the affordance planning process, rather than just in final outcomes**. While acceptable task results can improve user satisfaction and effectiveness, an understandable and adaptable planning process further improves communication efficiency by requiring less training data

for agent education [58, 78, 105, 115]. This ultimately leads to a better user experience and greater compatibility with humans.

Fulfilling the aforementioned goals is non-trivial. The first challenge is identifying relevant factors among various contexts and their collective impact on human affordance perception. Early studies modeled affordance understanding without context by detecting canonical affordance — the conventional use of an object — through vision [30, 76, 81]. To improve context awareness, later works, such as the SayCan series [3, 21, 47, 50, 120], inferred applicable actions with observations of the physical environment to ground the affordance knowledge of large language models (LLMs), others used more advanced vision language models (VLMs) to explore richer contextual semantics to improve the grounding [49, 104]. Additionally, several studies captured personal preferences for object locations in rearrangement tasks through propagation algorithms [59] and reinforcement learning with LMs [46, 85, 116, 118]. Although previous works gradually acknowledged contextual factors, they did not clearly identify the most important ones for object interaction planning and overlooked their collective impact, of which the complex interplay may influence how affordances are perceived and used in real-world scenarios. This issue limits the applicability of existing approaches in dynamic environments.

The second challenge stems from the demand for both desirable interaction results and an understandable, adaptable planning process to ensure each proposed move is reconcilable with human practices. This requires the planning to align with human cognitive reasoning structures. Recent research has combined structured reasoning, such as chain-of-thought, with VLMs to improve reasoning in complex tasks [71, 80]. While demonstrating the advantages of cognitive architecture for improving the success rates of the generated plans, these studies often overlook its potential of incorporating relevant contextual factors to enhance process understandability and adaptability.

To address the first challenge, we first conducted a formative study ($N = 14$) in a typical household scenario involving diverse agent-environment interactions to explore human perceptions and uses of real-world context in affordance-based planning. Guided by the 4E cognition theory [84] and our findings, we derived design requirements. We instantiated the physical, intrapersonal, and interpersonal context [24, 114] in a domestic environment as location and object occupancy, human status, and social relationship and owner-defined usage-rights, respectively.

To bridge the second research gap, upon the insights from the formative study, we propose a framework, *ACKnowledge*, that computationally models the dual-process cognition architecture [56] (Fig. 2) for intelligent agents to understand real-world context, reason about interaction candidates with their desirability, and propose acceptable plans in a human-comprehensible manner. Specifically, *ACKnowledge* simulates the intuitive thinking process (System 1 [56]) using a weighted knowledge graph (KG) of common physical entities and their interaction possibilities constructed based on existing crowd-sourced commonsense knowledge of affordance (ATOMIC2020 [52]). This allows agents to create theoretically feasible plans based on normative affordances given a high-level task description. Then, *ACKnowledge* exercises analytical thinking (System 2 [56]) to propose and rank alternative plans considering physical, intrapersonal, and interpersonal contexts. This is achieved by (1) automatically adapting affordance perception based on physical context via weight redistribution and object-node property inference in the KG and (2) constructing plans using a context-aware retrieval-augmented generation (RAG) approach. This component enables agents to generate acceptable planning results with an understandable process that allows user negotiation and error correction with the agent.

We evaluated our proposed approach through two studies. We first assessed the execution disambiguity and human acceptability of the interaction plans generated by *ACKnowledge* in a simulated household setting. We simulated how agents visually observe the environment via a camera and ran *ACKnowledge* and two baselines to generate detailed plans to carry out given tasks accordingly. The plans were presented both as machine-executable commands and in human natural languages. We measured the execution disambiguity rate in each condition and invited 21 participants to rate the plans on potential success, physical validity, usage-rights respectfulness, social appropriateness, and overall acceptability. Results indicated that *ACKnowledge*'s plans were unambiguous and specific enough to act upon for 98.78% of tasks in simulation, outperforming the baselines by 14.63%. Participants rated *ACKnowledge*'s interaction plans significantly higher than those of the baselines, particularly in overall acceptability, usage-rights respectfulness, and social appropriateness.

To further explore *ACKnowledge*'s negotiation and personalization capabilities in an understandable planning process, we conducted a real-world case study ($N = 9$). A Wizard-of-Oz agent first demonstrated the planning process of *ACKnowledge* in augmented reality to participants. After understanding the process, participants instructed the agent to correct any suboptimal plans. They then reconfigured the physical environment and personalized the household using an interactive web app to define usage-rights and affordance usage. The result shows that participants can interpret the *ACKnowledge*'s planning process and successfully correct unsatisfactory plans to acceptable ones through negotiation. They are also satisfied with the household personalization process and *ACKnowledge*'s corresponding planning results.

In summary, the contribution of this paper to the HCI community is threefold.

- We identify factors in physical, intrapersonal, and interpersonal contexts that are critical to human-compatible agent-environment interactions.
- We propose the *ACKnowledge*, a context-aware computational framework that is self-adaptive and user-adaptable to support human-compatible, affordance-based interaction planning for agents working in dynamic human environments.
- We validate the performance of *ACKnowledge* in terms of acceptable planning results and an understandable planning process through a simulation study and a case study in real-world settings, taking household tasks as an example.

## 2 Related Work

### 2.1 Affordance in the Physical World

Affordance, first proposed by Gibson and appropriated by Norman, refers to the possible interactions offered by the objects in an environment [41, 86]. Humans' perception of affordance is crucial

for guiding how their subjective brain and body interact with the real world. Affordance is widely viewed as "*relational*" [20, 26], leading to various theories on its perception formation. Shotter and Noble proposed that affordances emerge from practical experience [103], while some scholars model acquisition as a reinforcement learning process [70, 81]. In contrast, Costall et al. [26] argued that man-made artifacts' affordances are more closely tied to embodied human intentions than past experiences, such as chairs designed primarily for sitting but also allowing standing.

Regardless of the formation, some affordances (a.k.a. canonical or typical affordances) are naturally favored over other possible affordances [26]. This implies distinguishable variance in relation strength, which can be computationally modeled as the weight on object-affordance relations [88]. Another relativeness of affordance is with the context where an object is presented in [95]. The 4E (embodied, embedded, enactive, and extended) cognition theory [84] suggests that humans dynamically adapt their perception of affordance to the interdependent and dynamic real-world contexts to guide their interactive behaviors with objects. Researchers recognized such real-world contexts as integrating physical (physical environment), intrapersonal (e.g., personal identities, active status, etc.), and interpersonal (e.g., social relationships, etc.) contexts [24, 114]. This suggests that contextual factors may alter the weights of the object-affordance relations mentioned above [88]. Inspired by the relational account of affordance, our work proposes to simulate humans' perception of the object-affordance network as a weighted knowledge graph and adapt it to real-world contexts in specific tasks.

## 2.2 Physical Affordance in HCI

Real-world object affordance, or physical affordance [57], has been extensively studied in human-computer interaction (HCI). One research direction emphasizes utilizing objects' current affordances. Research like [43, 90, 117] focuses on translating physical affordances, such as grasping, into gestural commands, while InteractionAdapt [22] emphasizes context-sensitivity of such translation. Other works have explored reusing real-world objects based on their physical affordances in Virtual or Augmented Reality (VR/AR) as props [72] or as Tangible User Interfaces (TUI) [32, 54] to create embodied experiences. Research like VR Haptics at Home [36] has even investigated repurposing objects based on their real-world affordances in VR/AR. Another line of research focuses on creating new affordances. Some research investigates how virtual representations suggest possible affordances [35] or apply affordance knowledge to assist object retrieval in VR [55, 121]. Beyond VR/AR, HCI researchers have developed tangible devices or robots that adapt their perceived affordances to user needs [37, 82]. Additionally, some studies have addressed the formation of affordances, with Liao et al. [70] viewing it as a reinforcement learning process and Munguia et al. [81] using RL for precise object manipulation in human-robot interaction. While these studies acknowledge that object affordances influence human interaction methods, they primarily focus on the canonical affordances of existing objects or potential affordances for new creations. The perception of potential affordances with varied possibilities for everyday objects remains an area for further exploration.

## 2.3 Affordance-based Real-world Interaction Planning Methods

The fields of machine learning (ML), computer vision (CV), robotics, and language technologies (LT) have longstanding interests in affordance-based real-world interaction planning. The rise of embodied AI further attracts research attention to this topic [80].

*2.3.1 Computer Vision-based Affordance Detection.* Many works have achieved high accuracy in affordance detection with various CV techniques to determine the best object to complete a given task. For example, Bahl et al. trained a visual affordance model that estimates human-object interactions within a scene [10]. Large-scale datasets like Partnet [79], which provides fine-grained part-level affordance annotations, and AffordanceNet [29], with 23k shapes labeled by 18 affordance categories, support continuous improvements in CV models. However, existing datasets and models are usually restricted to limited affordances, focused on regional affordance detection, and only allowed static inputs. They cannot sufficiently accommodate high-level interaction planning requests in real-world tasks with diverse contexts.

*2.3.2 Large Language Model with Real-world Grounding of Physical Context.* To compensate for the lack of real-world usability and context awareness in CV-based works, recent studies proposed to integrate LLMs with visual real-world grounding [3, 21, 47, 50, 120]. SayCan [3], as an example, achieved a planning success rate of 84% and execution success rate of 74% with PaLM in the 101 robotic tasks in a real kitchen. More works followed SayCan, extending to tasks with an open vocabulary of object affordances [21, 50, 120] and an improved planning policy [47]. Multimodal grounding further enhances the depiction of real-world contexts [31, 65, 125]. PaLM-E introduced embodied language models integrating sensor modalities with language models to connect words and percepts [31]. Li et al. developed a model that demonstrated robust multimodal perception and reasoning capabilities, effectively following human intent while exhibiting adeptness in context learning [65]. While effective, these real-world grounding methods attended primarily to the physical context of affordance-based tasks to optimize the real-world applicability of LLM planning.

*2.3.3 Language Model-Based Interaction Planning Beyond Physical Context.* In reality, the interaction context of an object consists of its surrounding physical environment, personal preference of users, as well as social and cultural norms [23, 68]. In light of personal preference, works like [19, 46, 67, 85, 116, 118] applied learning-based approaches to educating household agents to align with humans' object arrangement preferences. Some other researchers considered social contexts and studied human-aware (e.g., status-aware, activity-aware, etc.) interaction planning for robots [4, 25, 44, 73]. Recent advancements in VLMs enable researchers to capture and understand the visual semantics of a scene and reason with commonsense knowledge. Vila [49] and MOO [104] excelled at robot action and object manipulation planning with VLM in interactive tasks. Researchers have attempted to employ cognitive architecture in the language model for more comprehensive and reliable reasoning to facilitate affordance planning in increasingly complex tasks. EmbodiedGPT [80] leveraged VLM with chain-of-thoughts (CoT)

reasoning to plan robotic tasks in the Franka Kitchen and Meta-World Benchmark datasets and achieved 1.6 times and 1.3 times performance improvements compared to models without CoT.

In summary, existing works on affordance-based interaction planning incorporated common sense and real-world contexts but did not align them with the beliefs and perceptions of individual users. While capable of successfully executing low-level commands and high-level plans, these methods lacked the flexibility to accommodate on-demand personalization and dynamicity in the planning process and results. There is a pressing demand for an adaptive and adaptable architecture to plan human-compatible object interactions in changing human environments by comprehensively considering physical and other human-related contexts.

### 2.4 Computational Cognitive Architecture in Real-world LLM-based Interaction Planning

Extensive discussions and explorations have been about applying computational cognitive architecture to LLM-based interaction planning [60, 106, 107]. Psychologists introduced the dual-process theory to describe human decision-making [56, 113], dividing cognition into intuitive thinking (System 1) and analytical reasoning (System 2). This theory highlights that both information retrieval and analysis contribute to effective interaction decisions. Inspired by such a cognitive structure, SwiftSage [71] computationally modeled the two systems using a small fine-tuned language model and an LLM to simulate fast and slow thinking, respectively. It surpassed other methods in 30 household interaction planning tasks from the ScienceWorld benchmark.

Some researchers proposed that grounding the perception of the environment with commonsense knowledge enables fast reasoning, representing intuitive thinking [102]. To model this commonsense-based perception system computationally, researchers have explored using knowledge graphs (KGs) [52, 53], which effectively serves LLM agents with external knowledge or auxiliary information [69, 92, 112, 126, 127]. KG-based grounding can enhance LLM's ability to reduce hallucinations and increase their understanding of and accordance with context. As for modeling the human analytical reasoning system, prior works experimented with a wide variety of methods besides direct application of LLM, including but not limited to reinforcement learning [11], deep learning with cognitive significance [60], and a combination of classical planning [8].

Inspired by the previous research efforts, we proposed adopting a KG of affordance common sense as the intuitive thinking system, updating and retrieving information from the KG based on real-world contexts, and simulating the analytical reasoning system through a reasoning chain. Our goal is to enable intelligent agents to plan object interactions in dynamic human environments, achieving human-acceptable results through an understandable process.

## 3 Study 1: Formative Study

Recognizing that humans' awareness of real-world contexts [6, 13, 95] and their cognitive architecture [113] are essential for effective interactions with objects in dynamic environments, we conducted a formative study to explore how humans manifest these abilities in actual households and their demands for agents working in their space on affordance-based interaction planning.

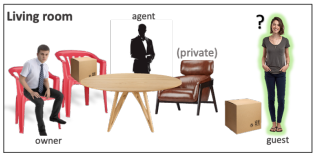
### 3.1 Study Design

The study included four semi-structured interviews where participants responded to two types of questions: one focused on their past experiences with interaction planning, and the other invited them to propose their expectations of the agents by imagining interacting with a fully capable agent in a household, alongside visual demonstrations as groundings (Fig. 1).

We chose a household scenario from various shared environments as it involves diverse objects, interactions, and social relationships. This choice also aligns with previous affordance-based planning works [3, 21, 47, 50, 120]. We designed the formative study in an imaginary format to avoid biases from humans' doubts about a robot's execution ability. This approach enabled us to explore human perceptions and usages of real-world contexts, together with their expectation of agents, in affordance-based interaction planning, free from the constraints of existing robot capabilities.

The first session explored participants' expectations and demands of intelligent housekeeping agents, helped them understand the background of intelligent housekeeping agents, and defined the scope of housekeeping tasks for the interview. Findings from this session guided our subsequent affordance-based interaction planning task designs (Section 5.1.1). The next three sessions investigated how humans perceive and utilize real-world contexts in affordance-based interaction planning across three settings with different context complexities: a general household, a personalized household, and a personalized household with multiple individuals. Each session featured a visual demonstration depicting specific room setup and task as interview probes (Fig. 1). Participants then answered interview questions from both human and agent perspectives based on their intuitive solutions and past experiences. This dual-perspective approach grounded in specific scenes and tasks allowed us to understand how users expect co-existing agents to perceive and utilize context comprehensively and concretely. Detailed interview questions are in supplementary materials.


**(a) The setting for the session of general household and personalized household. Only the owner and agent are in the living room/bedroom household. In the general household setting session, the task instruction is *"Find somewhere to place my bag"*. The room also serves as a hint for the personalized household setting session to inspire participants.**




**(b) The living room in the personalized household with multiple individual sessions. The scene now includes the owner, guest, and the agent, with the task *"Find my friend somewhere to sit"*.**

**Figure 1: Room settings and task scenarios presented to participants in semi-structured interviews.**

## 3.2 Participants and Procedure

The study received institutional IRB approval. We recruited 14 participants (7 males, 7 females; $Age = 23.14 \pm 2.03$, denoted as P1-P14) through social networks and word-of-mouth, four of whom (P1, P2, P4, P8) had prior experience with intelligent household agents. In the briefing, we explained the study structure and the concept of real-world contexts. We then discussed participants' expectations of intelligent housekeeping agents and their perceptions of contexts in various environments, session by session. Each interview lasted one hour (compensated $9) and was audio-recorded with the participant's consent.

## 3.3 Findings

Two authors conduct thematic analysis [14] on the interview transcripts through iterative coding and discussions and extract four key themes from the experts' responses: demanded assistance, real-world context factors, adaptable perception of affordance, and transparent planning process.

*3.3.1 F1: "Help me when I need."* We first learned that humans do not expect the agents to do everything for them. As P8 explained, *"I would like them to assist me with boring tasks or play the role of me when I am not home."* Based on our summarized codes, people generally seek agent assistance with specific household task types:

- Kitchen: Cleaning, Cooking
- Living Room: Arranging, Cleaning, Socialising
- Bedroom: Arranging, Cleaning, Entertaining, Assisting in Tasks in other rooms

*3.3.2 F2: "You should be considerate."* All participants, as homeowners, expected the agent to complete tasks according to real-world contexts and pointed out the specific contextual factors to consider when speaking from an agent's perspective. We adapted the concepts of physical, intrapersonal, and interpersonal contexts from [24, 114] and classified the factors mentioned in participants' feedback accordingly.

**Physical Context: Location, Occupancy**. Physical context refers to the physical environment where interaction takes place. Location significantly influences participants' interaction choices. For instance, all participants indicated that they would alter their planning if instructions were given in the bedroom instead of the living room. P6 reflected on a possible reason, *"Maybe some furniture has a higher hygiene standard [in the bedroom], and thus no longer suitable as a candidate to place my bag on."* This finding is an example of how characteristics of locations lead to different behaviors and norms of humans [2] in the form of the usage of objects. Moreover, seven participants highly expected the agents to detect the *"free area or items"*. Based on feedback, we identified critical occupancy-related properties for agents to consider: occupation (what affordance function they are currently offering), occupant (object/person that is using the object), and occupancy status (whether they are fully, partially, or not occupied). This occupancy concept helps computably model the situated affordance [84], which underlies how humans actually assess the feasibility and physical effort of using objects.

**Intrapersonal Context: Human Status**. Intrapersonal context involves the physical and mental status of an actor. Human awareness is crucial for agents. Nine participants highlighted the need for agents to recognize people's active status (i.e., static/dynamic) and posture (i.e., sitting/standing/lying), aligning with prior research [7]. This is because people's status could indicate their present interruptability and an acceptable level of engagement in upcoming interactions. P3 exemplified *"It would be annoying if the agent performed any noise-making action while I sleep"*. Outside the household, it is also usually inappropriate to interrupt your colleagues to help you when they are walking around distributing documents.

**Interpersonal Context: Social Relationship, Owner-Defined usage-rights**. The interpersonal context concerns the relationship between an actor and the other actors (and/or their belongings) in a shared environment. Regarding what the agent should consider in a multi-person environment, paying attention to social etiquette and relationships between individuals were mentioned six and five times. As social relationships are formed in various settings beyond households [18], relationship-awareness is also important in other shared environments. When discussing what rules non-owner users should obey when interacting with their households, eight participants respected private ownership and defined usage boundaries for their items. For example, P6 mentioned, *"Some utensils are specifically reserved for guests, and those I use personally are not freely available to others."* We concluded such personal preferences as owner-defined usage-rights and categorized them into three levels according to responses: private (exclusive to the homeowner), semi-private (restricted to the owner and designated users), and public. For example, a distant visitor may need permission to use the owner's private objects, whereas a close friend can use them without asking. Ownership, privacy, and personalization extend beyond households [9, 15]; individuals also own corners in offices or bunk beds in hostels, defining their own rules and expecting others to respect them.

*3.3.3 F3: "You should know me and my home."* Besides having some commonsense knowledge of the interaction possibilities, our participants expressed other requirements for agents, especially the need for agents to be adaptive. That is, the agents should be able to recognize the environment independently and reason about object affordances based on the existing environment. Eleven participants agreed that reasoning about an object's function based on its current state could enhance the interaction experience and save the owner's time educating the agent. Moreover, they all expected the agents' adaptive understanding of affordance to take their usage habits and item preferences into account in daily tasks. Ten participants stressed the importance of agent obedience and wished the agent to automatically learn their habits with minimal intervention and education. As P2 pointed out, *"Less is more, learn from observation."* Being adaptive is also important for agents outside the home, as they need to respond to dynamic environments and varying user preferences in public spaces.

*3.3.4 F4: "We should think on the same page."* While explainability [109] and personalizability [64] are recognized as vital for human-robot interactions, participants in our study emphasized that the agents should be understandable and adaptable by users, with adaptations transferable to future uses. They hoped to comprehend the

agents' decisions and behaviors and to personalize them through guidance, feedback, or corrections. Specifically, the participants expected the agent to know the *Do*s and *Don'ts*. *"They should have a sense of what they can do and cannot,"* stated P8. When inappropriate interactions with objects happened, some participants (P8, P9) preferred providing one-shot corrections for the agent to learn from its mistakes automatically. Other participants, instead, proposed to *"solve the problem from the bottom"* (P2); they imagined that such errors occurred due to the flaws in the agents' reasoning structure and would like to make direct adjustments to it. P11 said, *"If I can know what my agent is thinking and debug it, I might feel secure to assign more tasks to them."* Participants' opinions reflected the demand for transparency and modifiability in the agent's planning process to ease understanding and communication.

## 3.4 Design Requirements

Inspired by these findings, we derived four design requirements concerning what human users value and demand:

**F2, F3, F4 → DR1:** When planning, the agent should integrate affordance common sense with real-world context.

**F2 → DR2:** The agent should plan affordance-based interactions according to specific contextual factors.

**F3, F4 → DR3:** The agent's perception of affordance should be automatically adaptive to the physical environment and adaptable by the user.

**F3, F4 → DR4:** The agent should be able to generate acceptable plans with understandable and trustworthy process through user-friendly interactive modalities.

Based on the design requirements and the instantiated contextual factors from F2, we designed and implemented a housekeeping agent framework to plan affordance-based interactions compatible with humans.

# 4 Design and Implementation

In line with the DRs of the formative study, we developed the *ACKnowledge* framework integrating commonsense knowledge with awareness of contextual factors discussed in Section 3.3.2 as shown in Fig. 2. Comprising *"brain"* and *"vision"* modules that capture and synthesize real-world information in a dual-process approach, *ACKnowledge* identifies objects that can most effectively provide a specific affordance.

This section first introduces the primary *"brain"* and *"vision"* modules based on the weighted KG and vision language model (VLM). Section 4.2 explains how the *"brain"* and the *"vision"* modules coordinate in affordance-based interaction planning using real-world context. In Section 4.3, we demonstrated *ACKnowledge*'s interaction with users using a graphical user interface (GUI) for personalization and a conversational user interface (CUI) for error correction. Finally, we presented a user scenario to demonstrate the real-world usage of *ACKnowledge* regarding its configuration, resolution, and error correction.

## 4.1 Basic Modules of the *ACKnowledge* Framework

*4.1.1 "Brain" Module: Affordance Commonsense Scene Graph.*

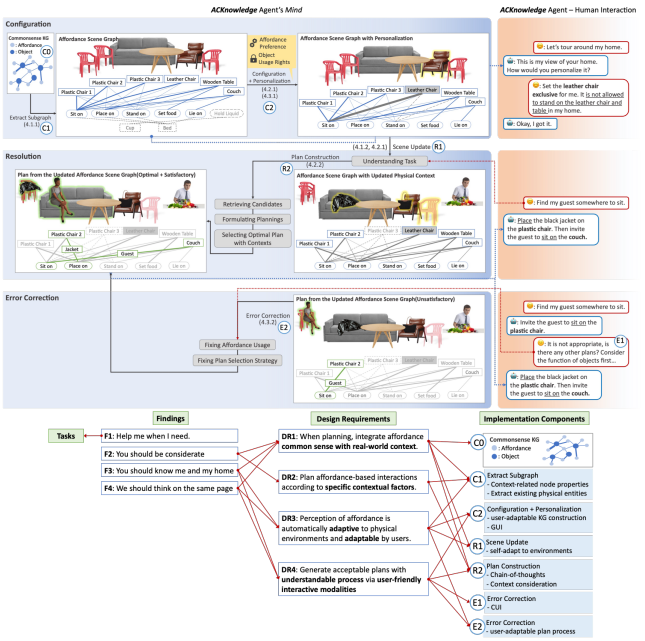



**Figure 2: The overall structure of *ACKnowledge*. Above figure: the left shows the perception and reasoning process of *ACKnowledge*, and the right shows the user-agent interactions. The workflow of *ACKnowledge* upon receiving instruction has three major phases: Configuration (C0-C2), Resolution (R1, R2), and Error Correction (E1, E2). This workflow also matches the user scenario as described in Section 4.4. Moreover, elements of the workflow model the dual process cognition. Step C1 is System 1 in the dual process, while Step C2, R1, R2, and E2 belong to System 2. The workflow of *ACKnowledge* has three major phases: Configuration, Resolution, and Error Correction. The below figure demonstrates how DRs are derived from findings and how each step reflects the DRs.**

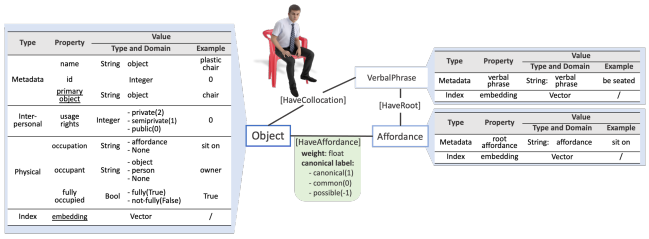



**Figure 3: The data structure in the affordance commonsense scene graph. The scene graph is the subgraph of the weighted affordance commonsense KG built from the ATOMIC2020 dataset [52] with augmented node properties. The underlined properties (primary object and embedding) are the initial properties of the object node in the original KG. The properties of VerbalPhrase and Affordance nodes remain the same in the scene graph and the original graph.**

**Weighted Affordance Commonsense Knowledge Graph.** Complying with the demand for commonsense knowledge (DR1) and psychology theories [88], *ACKnowledge* employs a weighted

KG to computationally model the connected and ranked relationship between physical entities and their affordance. We built this affordance commonsense KG upon crowdsourcing data from the ATOMIC2020 dataset [52]. It includes 2.8k tuples of affordance relationships between 62 common household objects and 1.5k affordances with 2.1k verbal phrases of affordance (Fig. 3).

Fig. 4 demonstrates the construction process in detail. We first selected the 62 commonly used household objects that occurred more frequently than average and obtained 11k related tuples from the *ObjectUse* (affordance relationship) relation tuples. Using spaCy, NLTK, and community detection from the SentenceTransformers package, we extracted and clustered verbal phrases from the original sentences into root affordances.

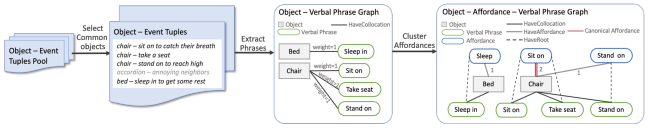



**Figure 4: Construction process of affordance commonsense knowledge graph. It follows the steps of selecting common daily objects, extracting related phrases to those objects, and finally, clustering extracted verbal phrases into affordances.**

Using three types of nodes — *Object* (62), *Affordance* (1.5k), and *VerbalPhrase* (2.1k) — we created three relations: *HaveAffordance* (2.8k), *HaveRoot* (2.8k), and *HaveCollocation* (4.0k). Each relation has a weight based on tuple frequency, indicating connection strength. We used frequency as weight because ATOMIC2020 balanced the appearance of each object and affordance. To label canonical (most typical), common (less typical), and possible (rare, occurred once or twice in the dataset) usages, we performed K-means clustering ($k = 3$) on *HaveAffordance* relations by weight, with higher weights indicating more canonical relations.

**Affordance Knowledge Scene Graph.** Based on the grand affordance commonsense KG, *ACKnowledge* extracts the sub-graph of existing physical entities in the real world as the localized scene graph (DR3). *ACKnowledge* also augments object node properties — such as owner-defined usage-rights and occupancy — to sufficiently describe the context (DR2). The usage-rights property of the *Object* nodes, the naming of the *Affordance* nodes, and the existence and weight of *HaveAffordance* relations are all modifiable by users (DR3). However, occupancy properties remain fixed as part of the ACKnowledge *"vision"* module performance. The complete list of node and relationship properties is shown in Fig. 3.

#### 4.1.2 *"Vision" Module: Vision Language Model Based Object Detection.*

The *"vision"* module simulates how the agent visually observes the environment. As shown in Fig. 5, such module of *ACKnowledge* is realized by capturing real-time video and feeding it into a VLM with a chain-of-thought (GPT-4o-20240513 supported by Microsoft Azure, no pretraining, temperature = 0.0, all VLMs below refer to the same setting if not specified) to obtain the amount and occupancy of physical entities and human status.

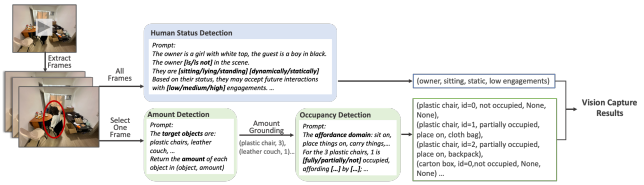



**Figure 5: The *"vision"* module process. It starts by extracting frames from a video and then sequentially detecting objects, their occupancy, and human statuses from the extracted frames.**

**Capturing the Amount and Occupancy of the Physical Entities.** *ACKnowledge* triggers physical entity capture when it constructs the scene graph and updates the environment for each task (DR2, DR3). With desired object entities as input, the VLM first detects the quantities. During the scene graph construction, all 62 objects are the desired objects, whereas, in each task, only task-related objects are detected for efficiency. The VLM then detects the occupancy accordingly by filling in the blank with the amount detection results as grounding.

**Capturing and Comprehend Human Status.** Human status, as a contextual factor required by DR2, is captured and comprehended using video input. The video input is first processed into frames with 5 fps, which prompts the VLM to describe the humans' active status (static or dynamic) and posture (standing, sitting, or lying) of each existing registered human character. VLM also speculates an acceptable engagement level (the human can accept which disruptive level of incoming interaction) for future interactions based on the present human status.

## 4.2 Cooperation of the *"brain"* and the *"vision"* with Real-world Context

*4.2.1 Automatic Affordance Perception Adaption.* *ACKnowledge* adapts its affordance perception according to the physical environment automatically through weight redistribution and node properties inference (DR3) by observing real-world context factors instantiated (DR2).

**Weight Redistribution Based on Location.** *ACKnowledge* expresses location-sensitive affordance preferences alternation by redistributing the weight according to the location using KG-LLM combined scoring. LLMs predict the probability of the occurrence of a specific word over the existing sequence $p(w_k|w_{<k})$ as *log-probs*. To integrate LLM knowledge with validated commonsense crowdsourcing knowledge, we formulated our goal as predicting $P(A_{obj}^{aff}|B_{loc}) = P(A_{obj}^{aff})P(B_{loc}|A_{obj}^{aff})/P(B_{loc})$, where $A_{obj}^{aff}$ denotes the event that one object *obj* affords the affordance *aff*, and $B_{loc}$ denotes the event's location is *loc*.

$$
\begin{aligned}
&P(A_{obj}^{aff}) = \frac{weight(aff, obj)}{frequency(obj, KG)} \\
&P(B_{loc}) = 1/3, \text{assume the probability of choosing one room out of three is equal} \\
&P(B_{loc}|A_{obj}^{aff}) = e^{logprob(loc|A_{obj}^{aff})}
\end{aligned}
\tag{1}
$$

where *weight(aff, obj)* denotes the weight on the *HaveAffordance* relation between *obj* and *aff*, *frequency(obj, KG)* denotes the frequency of *obj* in the whole *KG*. *logprobs* is the prediction of text-generation probability by LLM (text-embedding-002, no pretraining, temperature = 0.0). Finally, *ACKnowledge* employs the normalized weights to redefine affordance perception for particular locations.

**Weight Redistribution Based on Occupancy.** *ACKnowledge* redistributes the weights of KG after updating the occupancy factor to reflect the occupancy's cue for preference. The weight of the affordance that the object currently affords upgrades to the maximum weight of all, while others adjust proportionally to maintain a consistent total weight.

**Usage-rights Property Inference.** In addition to weight distribution, *ACKnowledge* infers usage-rights properties based on vision. The appearance and number of objects hint at usage-rights. Three plastic chairs will likely be public, whereas one delicate leather chair will likely be owner-exclusive. *ACKnowledge* calculates the variance in the number and clusters of objects belonging to the same primary object category. If the variance is equal to or greater than 1.0, K-means clustering ($k = 2$) is triggered to distinguish a unique object and mark it as private in terms of usage-rights.

*4.2.2 Affordance Interaction Planning Construction chain.* *ACKnowledge* also uses the context information in the interaction-planning reasoning process with an LLM-based retrieval augmented generation method. Such a RAG-supported reasoning structure contributes to the process understandability (DR4). As shown in Fig. 6, the process compromises four consecutive steps: understanding the task, retrieving candidates with real-time environmental updates, formulating planning candidates, and selecting an optimal plan based on real-world contexts.

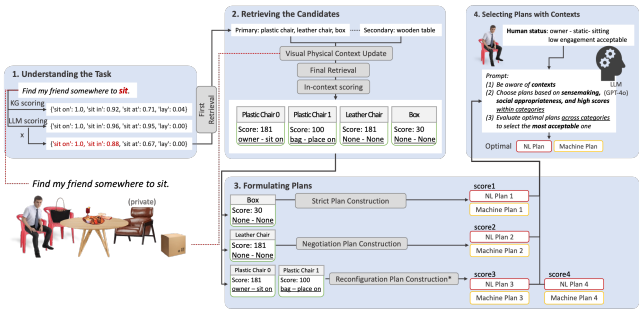



**Figure 6: The process of plan construction consists of four steps: understanding the task, retrieving the candidates, formulating the plans, and selecting optimal plans with contexts. *ACKnowledge* provides three candidate plans for each task: Strict, Negotiation, and Reconfiguration, explained below *Step 3. NL Plan* denotes human-friendly natural language plan, *Machine Plan* denotes machine execution plan.**

**Step 1: Understanding the Task.** When receiving an *instruction*, *ACKnowledge* uses the LLM (GPT-4o-20240513 supported by Microsoft Azure, no pretraining, temperature = 0.0, all LLMs below except for logprobs purpose refer to this same setting) to extract the key affordances, $k_1, k_2, \ldots, k_n$, with corresponding occupants, $occ_1, occ_2, \ldots, occ_n$, implied in the text. Using key affordances extracted by the LLM as input, for each key affordance $k_i$, *ACKnowledge* retrieves a list of weighted KG domain affordances, denoted as $Aff_i^{(1)}, Aff_i^{(2)}, \ldots, Aff_i^{(n_i)}$, with normalized similarity scores based on *VerbalPhrase* and *Affordance* node embeddings, denoted as $\{\text{KGSimilarity}(k_i, Aff_i^{(j)}) | j = 1, 2, \ldots n_i\}$. To calculate how accurately can an affordance $Aff_i^{(a_j)}$ express the keyword $k_i$, *ACKnowledge* calculate the $\text{KeywordSimilarity}(k_i, Aff_i^{(a_j)}) = \text{KGSimilarity}(k_i, Aff_i^{(a_j)}) \times \text{LLMSimilarity}(k_i, Aff_i^{(a_j)})$, combining the KG similarity scoring with an logprobs-based LLM (GPT-3.5-turbo-instruct, no pretraining, temperature = 0.0) similarity scoring like in Section 4.2.1. Finally, *ACKnowledge* selects the most related key affordances candidates using K-Means clustering ($k = 2$).

**Step 2: Retrieving Candidates with Real-time Environment Updates.** *ACKnowledge* retrieves primary objects, which are the ones that are directly linked to the selected affordances in the previous step, and secondary objects, which are less than four steps away from primary objects. Subsequently, *ACKnowledge* calls the selective occupancy detection (4.1.2) only for primary and secondary objects to save computing resources. Having obtained those occupancy updates, *ACKnowledge* redistributes weights (4.2.1) in the *"brain"* module. With the updated KG in the *"brain"* module, *ACKnowledge* returns a set of possible candidates linked to the affordance candidates of the extracted key affordances, each with a score. For each object candidate $obj_i^{(j)}$ that can afford the key affordance $k_i$ with $occ_i$ as occupant, *ACKnowledge* calculates its CandidateScore, a non-zero score indicates the target $obj_i^{(j)}$ is vacant to afford $k_i$ with $occ_i$ as occupant:

$$\begin{aligned}&\text{CandidateScore}(k_i, occ_i, obj_i^{(j)}) = \\ &\text{AffordableScore}(obj_i^{(j)}) \times \text{OccupancyScore}(obj_i^{(j)}, occ_i)\end{aligned} \tag{2}$$

where

$$\begin{aligned}&\text{AffordableScore}(\text{obj}_\text{i}^{(\text{j})}) = \\ &\sum_{x=1}^{n_i} weight(Aff_i^{(x)}, obj_i^{(j)}) \times \text{KeywordSimilarity}(\text{k}_\text{i}, \text{Aff}_\text{i}^{(\text{x})}) \\ &\text{OccupancyScore}(\text{obj}, \text{occ}) = \\ &\text{ObjectSimilarity}(\text{obj.occupant}, \text{occ}) \\ &\times \text{OccupancyPotential}(\text{obj.fully\_occupied})\end{aligned} \tag{3}$$

The $\text{ObjectSimilarity}(obj.occupant, occ)$ is embeddings' cosine similarity; The $\text{OccupancyPotential}(obj.fully\_occupied)$ equals to 1 if $obj.fully\_occupied$ is False, otherwise 0.

**Step 3: Formulating Planning Candidates.** *ACKnowledge* employs an object-based strategy for plan formulation, as depicted in Fig. 6 *Formulating Plans* module, outputting natural language plans and machine execution plans. Based on usage-rights and vacancy (indicated by the CandidateScore calculated previously), *ACKnowledge* classifies candidates into three types and then formulates and evaluates strict, negotiation, and reconfiguration plans.

- Strict Plans: Directly use the usage-rights-obedient and vacant target candidates. Plan scores equal to the CandidateScore.
- Negotiation Plans: Ask for the house owner's permission, then use the usage-rights-disobedient yet vacant target candidates. Plan scores equal to the CandidateScore.
- Reconfiguration Plans: *ACKnowledge* uses a recursive reconfiguration method to vacate non-vacant targets by finding substitute objects for the current affordances of the targets (see supplementary material Algorithm 1). If substitutes are occupied, *ACKnowledge* deepens the search and recursively creates a reconfiguration plan until a vacant substitute is found or the search depth reaches the threshold (set to two). It then retraces steps to generate the complete reconfiguration plan, with plan scores calculated as the product of the CandidateScore of the final object-affordance-occupant tuples returned by the algorithm.

Strict plans are optimal solutions based on fixed real-world contexts, while negotiation and reconfiguration plans adapt the physical or interpersonal context for optimal outcomes. We maintained the intrapersonal context, particularly the human status factor, to minimize disruption to ongoing activities.

After constructing plans in lists of object-affordance tuples, *ACKnowledge* composes machine execution plans with four basic commands (pick-up, go-to, place, request) and the final target affordance command by rule-based text generation. We also utilized the LLM to generate human-friendly natural language plans.

**Step 4: Selecting an Optimal Plan Based on Real-world Context.** To incorporate intrapersonal context into optimal plan selection, we input plan candidates, human status descriptions (from Section 4.1.2), and selection guidelines into the LLM. Leveraging LLM's natural-language reasoning, *ACKnowledge* directs the LLM to prioritize plans based on sensemaking, minimal disruptiveness, and high category scores, then selects the most suitable one across categories. The LLM outputs the optimal plan and stores alternatives for future reference.

## 4.3 Interaction between *ACKnowledge* and the Users

In real usage, humans interact with *ACKnowledge* in three phases: configuration, resolution, and error correction, where configuration happens once while resolution and error correction occur on demand. Here, we introduce the user interface of the configuration and error correction phase.

*4.3.1 GUI for Personalization in Configuration Phase.* We developed a web app GUI for personalization (Fig. 7). After the room scans (photos taken during the room tour guided by the user) are uploaded, users can correct KG construction errors (e.g., amounts) and configure their household preferences. Based on pilot study results indicating that users preferred adjusting affordance properties related to specific objects, we organized functionalities on the homepage and object pages for better usability. The homepage displays a visual configuration of the household with an interactive KG and a dropdown menu listing all household objects and their default predicted usage-rights properties (Section 4.2.1). Users can

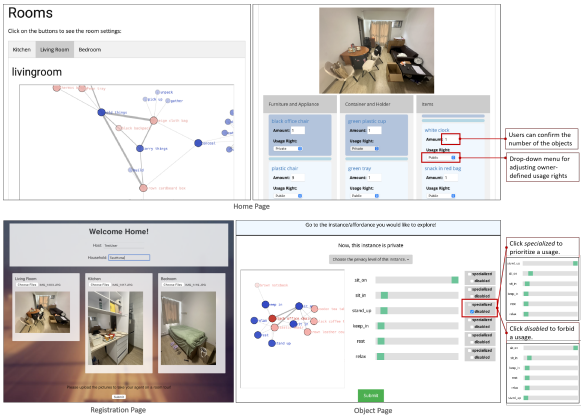



**Figure 7: GUI for user personalization during the *Configuration Phase*, including the registration, home, and object page. Users register their household by confirming the upload of the room scan pictures on the registration page. Then, the home page demonstrates the scene graph with pictures of the room, owner-defined usage-rights, and amounts of objects. Users can then adjust the owner-defined usage-rights and affordance usage of objects on the home and object pages.**

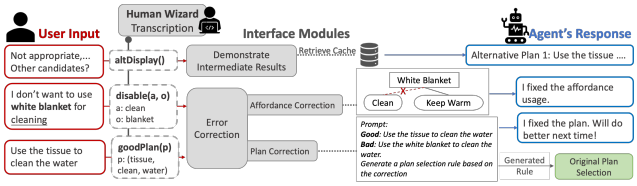



**Figure 8: CUI used in error correction process. Wizard-of-Oz assistance happens when transcribing and extracting keywords from users' speech. Users can interact with the *ACKnowledge* in the form of intermediate candidate demonstration, affordance-based error correction, and plan-selection-based error correction.**

customize their preferences for object usage-rights. By clicking on an object's name, users are directed to the object's page to adjust their preferences by prioritizing particular affordances (assigning the highest weight and redistributing the remaining proportionally) or disabling others (setting the weight to zero). Feedback from pilot users prompted us to offer checkbox options with timely result displays, which reduce confusion compared to scroll bars. All changes will then be updated to the corresponding *Object* nodes or *HaveAffordance* relations in KG.

*4.3.2 CUI for Error Correction.* During the error correction phase, users interact with *ACKnowledge* through a conversational interface (Fig. 8), with human wizard only employed for transcribing to avoid automatic speech recognition error affecting users' primary focus of the evaluation. The CUI supports two response types: intermediate candidate demonstration and error correction. The collaboration of the two response types helps users locate and fix the error more efficiently with less effort.

- Intermediate Candidate Demonstration: Upon queries about alternative objects or plans, *ACKnowledge* retrieves candidates from plan construction (Section 4.2.2) Step 1 or suggests alternative plans from the plan cache of Step 4.
- Error Correction: For affordance correction, users can disable an object's affordance realized by *ACKnowledge*'s setting the weight of a certain *HaveAffordance* relation in KG to zero. Regarding plan selection correction, if users specify a selection strategy, *ACKnowledge* integrates this new strategy into plan construction (Section 4.2.2) Step 4 (plan-with-mannual-rule correction). Otherwise, *ACKnowledge* generates a new selection strategy using LLM and inserts it into Step 4 (plan-with-automatic-rule correction). These two types of correction are distinctive and complementary so that users can easily opt for one strategy when an error happens, enhancing the usability of the CUI.

## 4.4 User Scenario of *ACKnowledge*

We describe a homeowner, Adam, who uses the intelligent housekeeping agent supported by *ACKnowledge* as shown in Fig. 2. During the configuration phase, they tour the agent around, let it scan the rooms and confirm the upload of its observation to the server (Fig. 2_C1). The system extracts a subgraph from its knowledge base (Fig. 2_C0) to create a scene graph, capturing default affordances and object usage-rights with Adam's personalization adjustments and correction of construction errors (Fig. 2_C2). When Adam's colleague Belle arrives for dinner, they instruct *ACKnowledge* to "Find my guest somewhere to sit" while they cook. The system scans the living room for available seating (Fig. 2_R1) and identifies empty chairs stacked in the corner while a jacket occupies the couch. It concludes that the couch is now used for storage and suggests the chairs for Belle (Fig. 2_R2). Unsatisfied, Adam asks *ACKnowledge* to *"move the jacket from the couch to the plastic chair and invite Belle to sit on the couch"* (Fig. 2_E1). The system accepts their suggestion, moves the jacket to the plastic chair, and invites Belle to sit on the couch. To improve future planning, *ACKnowledge* learns from Adam's correction and retains this selection strategy for future reference (Fig. 2_E2).

# 5 Study 2: Performance Evaluation in Simulations of Personalized Household

This section presents a two-factor within-subject study that assesses the execution disambiguity rate and user subjective feedback of an *ACKnowledge*-powered agent in a simulated household setting.

## 5.1 Experiment Design

In this experiment, participants took on the role of the house owner, Alice, in a simulated household with personalized configurations. They evaluated three intelligent housekeeping agents by comparing their proposed interaction plans' potential success rates and acceptability for different tasks.

### 5.1.1 *Tasks and Simulation Settings.*

**Simulation of Personalized Household.** We set up the simulated household in a campus dormitory with a living room, a kitchen, and a bedroom, which imply different social functions. To simulate

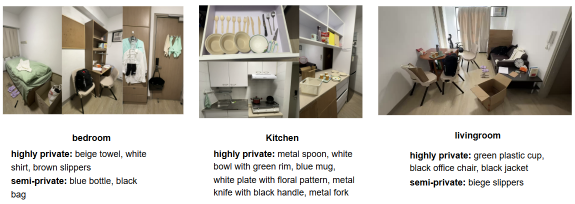



**Figure 9: The settings of three rooms, labeled with the usage-rights of objects.**

Alice's household, we add objects with distinguishable appearances from 62 everyday objects (Section 4.1.1) to each room. We also simulated Alice's personalization of the household by assigning different usage-rights to objects. Two human characters, owner Alice (who appears in all three rooms) and guest Bob (who appears in the living room and can only use public objects without Alice's permission), are involved. Fig. 9 shows the settings of three rooms with descriptions.

**Instructions and Tasks.** As shown in Fig. 10, we first selected


**Figure 10: The process of task construction. After extracting 50 common affordances, we clustered them into 13 affordance activity categories and assigned them to each room according to human desired tasks as investigated in 3.3.1. Finally, we constructed the 82 tasks with different environmental configurations.**

50 commonly used affordances (frequency above or equal to average) from the object-affordance tuples collected in Section 4.1.1. After confirming coverage of those in SayCan [3], we clustered them into 13 general activity categories using GPT-4o-20240513 (no pretraining, temperature = 0.0). Affordances were assigned to each room based on desired tasks from the formative study. Two participants, not involved in the study, created instructions for each room based on these affordances with categories, resulting in 82 instructions (27 for the bedroom and kitchen, 28 for the living room). We developed 82 tasks, with 20 using default room configurations and 66 featuring updated arrangements and human characters. We filmed the settings with two actors at fixed angles and evaluated the agent's machine execution plan alongside the human-friendly natural language plan (Section 4.2.2).

### 5.1.2 *ACKnowledge and Baselines.*

We compared the agent based on the *ACKnowledge* framework with two LM-based agent baselines to explore (1) Is context awareness a must to improve LM-based agents' performance? (2) Is computational thinking and reasoning architecture making interaction planning more effective?

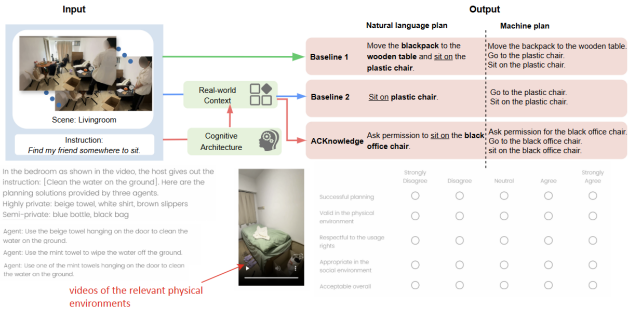



**Figure 11: Above figure: The demonstration of the structure of the two baselines compared with the *ACKnowledge* and the questionnaire. Below figure: A screenshot of the questionnaire. Users rate three solutions from the three agents based on textual descriptions and physical environments presented in video form in the questionnaire.**

All three agents first went through the configuration phase (introduced in Section 4.3), simulated by receiving pictures of rooms and house owner's personal configurations (usage-rights) as input. *ACKnowledge* constructed its own scene KGs upon receiving the input, while the baselines stored it in memory without incorporating KGs. Proceeding to the resolution phase, each agent then observed the updated physical environment (through pre-recorded videos) and generated plans accordingly. As shown in Fig. 11, the architectures are not the same across all approaches. In particular, Baseline 1 lacks context awareness and cognitive architecture, functioning solely as a VLM without contextual prompts or a Chain-of-Thought plan construction (i.e., no *understanding task -> retrieve object candidates -> form and select optimal plan* sequence as described in Sec. 4.2.2). In contrast, Baseline 2 incorporates contextual hints in its prompts but lacks cognitive architecture (e.g., "You should consider contextual factors such as object occupancy and usage rights ...") and does not follow the decomposed reasoning chain employed by *ACKnowledge*. They all produced the same format outputs: a machine execution plan and a human-friendly natural language plan.

#### 5.1.3 *Metrics.*

We evaluated the agents' performance with six metrics in Table 1.

**Table 1: Metrics for evaluating the agents' performance.**

| Type | Metric | Definition |
|---|---|---|
| Objective | Execution Disambiguity Rate | Can the plan be executed without ambiguity? |
| Subjective | Plan Successful | Can the plan meet the demands of the human's instruction? |
| | Physically Valid | Are the objects involved in the plan perceived functionally valid and vacant? |
| | Usage-Rights Respectful | Is the plan using objects respecting usage-rights defined by the owner? |
| | Socially Appropriate | Is the plan not burdening/disrupting the social beings and social environment (if any)? |
| | Overall Acceptable | Is the plan acceptable overall in the current context? |

The *execution disambiguity rate* and *plan successful* metrics are adapted from SayCan[3]. The *overall acceptable* metric evaluates general planning-result-oriented acceptability (DR4). The other three metrics are proposed to evaluate the awareness of the real-world contexts as instantiated in Section 3.3.2 and required by DR2.

To assess the execution disambiguity rate, we compared the physical entities automatically parsed from the agents' machine plans with those in the room configuration. If all entities are present (including the ones chosen to accomplish the instruction and others mentioned in the plan), the plan is execution-wise successful. For the other five metrics, we collected subjective ratings via questionnaires on a 5-point Likert scale (1 as the worst) from participants.

## 5.2 Participants and Procedure

With the approval of institutional IRB, we recruited 21 participants (11 male, 10 female, *Age* = 22.38 ± 1.96) of different majors and occupations from the campus through social media and word of mouth. None of them have participated in the formative study. Each participant spent two hours filling out the questionnaire (compensated $9 per hour).

Participants first got familiar with the background of the study in a 15-minute video briefing introducing the general tasks, room settings, and five subjective rating metrics. Afterward, participants started to rate the planning resolutions to the 82 tasks given by *ACKnowledge* and two baselines. For each task, participants were presented with videos of the physical environments where the plans were supposed to take place. The order of the resolutions from the three methods was shuffled. At last, we collected the ratings from the 21 participants, each with completion of all 82 tasks.

## 5.3 Results

This section presents the objective performance and subjective rating results of *ACKnowledge*, Baseline 1, and Baseline 2.

#### 5.3.1 *Objective Performance: Execution Disambiguity Rate.*

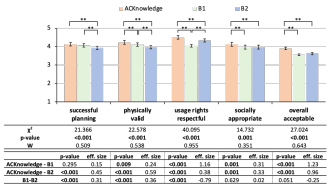
*ACKnowledge* achieved a high 98.78% disambiguity rate in its task execution with only one failure case caused by a wrong occupancy detection. By contrast, Baseline 1 and 2 succeeded in 84.15% and 80.49% of the cases. We conducted an error analysis and found that most errors happened because of ambiguities/hallucinations in the chosen objects, and three happened because of failing to generate a plan in five trials, as the generation may have violated the content policy. The outstanding performance of *ACKnowledge* proved the necessity of real-world grounding in visual context understanding and affordance reasoning.

#### 5.3.2 *Subjective Ratings.*

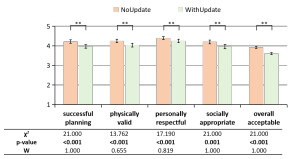
Since the ratings do not comply with normal distribution, we conducted a within-subject non-parametric Friedman test followed by post hoc Wilcoxon signed-rank tests and presented the results in Fig. 12.

**Overall subjective ratings.** Using the Friedman test, we compared the performance of *ACKnowledge* and the two baselines in terms of the five metrics introduced in Section 5.1.3 of two types of tasks (with/without environmental update). Both the method and the task type were found to affect the ratings significantly.

*ACKnowledge* received the highest ratings in all five metrics. The pairwise comparisons showed that *ACKnowledge* significantly outperformed Baseline 2 in all five metrics while significantly surpassing Baseline 1 regarding usage-rights respectfulness, social appropriateness, and overall acceptability. Noticeably, compared with the second-best methods, the overall acceptability of *ACKnowledge* (3.90 ± 0.07) significantly increased by 0.28. Moreover, usage-rights respectfulness, and social appropriateness increased significantly




(a) Subjective ratings of ACKnowledge and the two baselines across five metrics in all tasks.



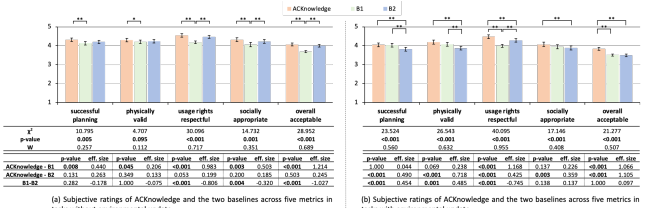

(b) Significance analysis between two task types.

**Figure 12: Subjective ratings in all tasks and significance analysis between two task types. The ratings are on a 5-point Likert scale (1 being the worst). Error bars depict standard errors. Friedman test was employed for more than two types of methods, and post hoc Wilcoxon signed-rank tests were used to assess specific pairwise differences. Significance values are reported for p < .05 (*) and p < .01 (**) after one-step Bonferroni correction, abbreviated by the number of stars. We calculated and presented Hedges g as effect size indicators for significant comparisons.**

by 0.16 and 0.15. Such improvements undermined that *ACKnowledge* has enhanced awareness of human-factors-related contexts, contributing to more acceptable interaction planning.

Baseline 1 outperformed Baseline 2 significantly regarding plan success and physical validity yet fell behind or showed no significant difference regarding usage-rights respectfulness and social appropriateness. Interestingly, both baselines were rated similarly regarding overall acceptability. This trade-off between metrics implies that metrics related to human sentiments, such as usage-rights respectfulness and social appropriateness, more strongly influenced perceived acceptability.

**Subjective ratings in different task types** Significant differences were found between the task types in terms of all five metrics. Thus, we further analyzed results within each task type (Fig. 13).




(a) Subjective ratings of ACKnowledge and the two baselines across five metrics in tasks without environmental update.

(b) Subjective ratings of ACKnowledge and the two baselines across five metrics in tasks with environmental update.

**Figure 13: Subjective ratings in different task types. The ratings are on a 5-point Likert scale (1 being the worst). Error bars depict standard errors. Friedman test was employed for more than two types of methods, and post hoc Wilcoxon signed-rank tests were used to assess specific pairwise differences. Significance values are reported for p < .05 (*) and p < .01 (**) after one-step Bonferroni correction, abbreviated by the number of stars. We calculated and presented Hedges g as effect size indicators for significant comparisons. B1 denotes Baseline 1 and B2 denotes Baseline 2.**

In tasks without environment updates, significant differences between the methods were found. *ACKnowledge* performed best overall (4.07 ± 0.07), with Baseline 2 as the second (4.00 ± 0.07). *ACKnowledge* rated all-round highest but found no significant difference with Baseline2 in all metrics in pairwise comparisons. In addition, there is no significant pairwise difference between Baseline1 and Baseline2 regarding plan success and physical validity. Thanks to the context awareness during both the configuration and the resolution process, Baseline2 performed as well as *ACKnowledge*. However, Baseline 1 did not perform well because context awareness was not noted in the resolution process.

In tasks with environment updates, there were also significant differences between the methods regarding all five metrics. *ACKnowledge* remained the best significantly regarding overall acceptability, but Baseline 1 became the second best with no significant favor shown over Baseline 2. In the other four metrics, no significant difference was found between *ACKnowledge* and Baseline 1 regarding plan success, physical validity, and social appropriateness. The result that Baseline 1 excelled over Baseline 2 while performing nearly as well as *ACKnowledge* requires attention. One reason is Baseline 1's lack of context awareness, which allowed it to exploit affordance possibilities more effectively. Being not sensitive to the predefined preferences and the updated object vacancy, it ignored some restrictions and found valid candidates to generate feasible plans. Another reason is that Baseline 2's strict obedience to context-based rules limited its reasoning capabilities. Baseline 2 struggled to propose alternative reconfiguration or negotiation plans without a cognitive reasoning architecture when the best candidate was not directly usable, unlike *ACKnowledge*.

## 6 Study 3: Exploring the Understandability and Usability of *ACKnowledge* in Real-world Settings

To explore the potential of *ACKnowledge*'s planning process to be understood and adapted by humans via usable interfaces, we conducted a case study inviting participants to interact with a Wizard-of-Oz agent built upon *ACKnowledge*.

### 6.1 Experiment Design

The study took place in a real home with a kitchen, living room, and bedroom, consisting of three phases: plan process demonstration, error correction, and personalization. In the demonstration and error correction phase, the household configurations were inherited from Alice's household in the simulation study, with participants acting as Alice. The *ACKnowledge*-based agent had knowledge of the environment settings and personalizations, as outlined in Section 5.1.1. In the final personalization phase, participants viewed the household as their own, allowing them to customize it while the agent completed tasks based on their configurations.

We measured understandability (DR4), trust (DR4), negotiability (DR3, DR4), and adaptability (DR3, DR4) in the planning process, along with the usability of interactive interfaces (DR4) through interviews. Additionally, we assessed planning outcomes after error correction and personalization using subjective ratings based on the metrics described in Section 5.1.3. Detailed tasks for each phase are available in the supplementary materials.

*6.1.1 Demonstration phase.*
In this phase, participants wore VR glasses to experience the interaction planning process of an intelligent agent in Augmented Reality. We selected three interaction planning tasks (one per room) where *ACKnowledge*'s solutions were rated as acceptable from the simulation study to demonstrate *ACKnowledge*'s planning process. After participants assigned tasks to the agent, the agent moved around and voiced its planning based on the intermediate result of *ACKnowledge*. Having watched demonstrations in all rooms, participants were interviewed to interpret the planning process in their own sense and provide feedback on its understandability and trustworthiness.

*6.1.2 Error-correction phase.*
After users interpreted how *ACKnowledge* worked, we investigated its negotiability with an error correction study. We selected three tasks (one per room) where *ACKnowledge*'s solutions were rated unsatisfactory regarding physical validity and/or social appropriateness in the simulation study (Section 5). For each task, we asked to rate the original plan solution using metrics from Section 5.1.3. If dissatisfied, they could communicate with the agent through the CUI (Section 4.3.2). Based on user feedback, the agent adapted its reasoning in real-time to propose better solutions (improved plan), which participants then rated. Finally, we created a similar task for *ACKnowledge* to solve and invited participants to rate it to test if the agent could automatically transfer the adaptation (transferred plan). We compared participants' ratings of the three plans to evaluate error correction outcomes and interviewed them about the usability of the CUI and negotiability in the error correction process.

*6.1.3 Personalization phase.*
In the final personalization phase, participants personalize their homes after arranging objects as desired. After participants guided the agent through the household and confirmed the upload of its room scans to the server (simulated by photos taken under their instruction), they used the personalization GUI (Section 4.3.1) to set their preferred usage-rights and affordances. An interview followed to gather their feedback on the adaptability of the personalization process and the GUI's usability. Participants then chose their appearances in the rooms. The agent observed the environment through real-time video capturing and planned interactions for the same three tasks from the demonstration phase. Participants evaluated these new interaction plans with ratings and assessed the process's adaptability and the alignment of outcomes with their personalizations through interviews.

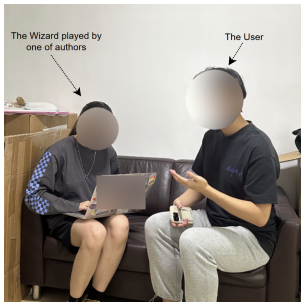
## 6.2 Participants and Procedure

Approved by institutional IRB, we recruited 9 participants (5 male, 4 female, *Age* = 22.56 ± 1.71, denoted as U1-U9) from the campus through social media and word of mouth. None of them participated in the formative study, ensuring they had no previous knowledge of the potential structure of *ACKnowledge*.

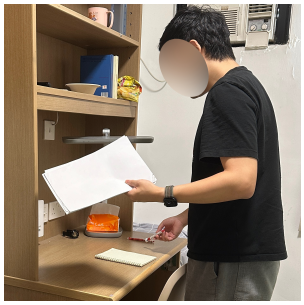
After a short briefing, the participants participated in the order of demonstration, error correction, and personalization phase (as shown in Fig. 14). They first wore VR glasses to interact with the AR agent during the demonstration phase to enhance their understanding of the agent's real-world usage. Once they grasped the process, participants completed the remaining phases without VR glasses. The study lasted 1.2 hours, with each participant compensated $12.


**(a) A User's first-person view through VR glasses during the demonstration phase.**

**(b) A user was in the demonstration phase.**




**(c) A user was in the error correction phase.**



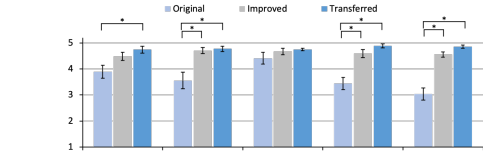
**(d) A user was in the personalization phase.**

**Figure 14: Actual scene photos from the case study.**

## 6.3 Results

*6.3.1 Demonstration Phase.* We conducted a thematic analysis on the participants' interpretation of *ACKnowledge*'s reasoning process, following the procedure in the formative study (3.3). The participants could describe *ACKnowledge*'s rationales in their words and recognized its planning process as being context-aware with evolving reasoning.

More specifically, the participants thought that ***ACKnowledge* captured physical, intrapersonal, and interpersonal contexts in its planning process**, acknowledging "the consideration of whether this object can execute such task" (U1-U9), "the possibilities provided by the environment" (U3, U6), "social factors that involved humans' activities, identities, etc." (U1, U3, U4), and "the application of usage-rights to filter out some candidates" (U2, U3). All participants (U1-U9) agreed that ***ACKnowledge* chose the most contextually acceptable plan based on its standards**. Everyone grasped the overall strategy employed by the agent. Four (U3, U6, U7, U9) noticed the process of locating key affordance implied by the task. Only one participant (U2) found the intermediate steps confusing due to his expectation of explanation for detailed reasoning. U6 mentioned the subjectivity of an "optimal" plan, indicating the need for *ACKnowledge* to be communicative and personalizable during planning. All participants concurred that **the planning process of *ACKnowledge* was reasonable**, some indicated that




| | successful planning | physically valid | usage rights respectful | socially appropriate | overall acceptable |
|---|---|---|---|---|---|
| $\chi^2$ | 13.188 | 13.310 | 2.100 | 15.800 | 16.222 |
| p-value | **0.001** | **0.001** | 0.350 | **<0.001** | **<0.001** |
| W | 0.733 | 0.739 | 0.117 | 0.878 | 0.901 |

| | p-value | eff. size | p-value | eff. size | p-value | eff. size | p-value | eff. size | p-value | eff. size |
|---|---|---|---|---|---|---|---|---|---|---|
| Original - Improved | 0.117 | -0.903 | **0.042** | -1.516 | 1.000 | -0.460 | **0.042** | -1.827 | **0.012** | -2.714 |
| Original - Transferred | **0.042** | -1.330 | **0.042** | -1.646 | 0.409 | -0.662 | **0.012** | -2.610 | **0.012** | -3.405 |
| Improved - Transferred | 0.093 | 0.584 | 1.000 | 0.219 | 0.829 | 0.249 | 0.293 | 0.766 | 0.082 | 1.181 |

**Figure 15: Subjective ratings of ACKnowledge in original plans, improved plans, and transferred plans on similar tasks. The ratings are on a 5-point Likert scale (1 being the worst). Error bars depict standard errors. Friedman test was employed for more than two types of methods, and post hoc Wilcoxon signed-rank tests were used to assess specific pairwise differences. Significance values are reported for p < .05 (*) and p < .01 (**) after one-step Bonferroni correction, abbreviated by the number of stars. We calculated and presented Hedges g as effect size indicators for significant comparisons.**

reasonability **enhanced the trust**. U3 and U9 found it "similar to human reasoning." while U1 and U7 echoed, "the clear planning process makes the agent seemingly more trustworthy and helpful." However, several noted that "*ACKnowledge* overthinks than humans" (U1, U4, U6). They argued that when selecting candidates, humans relied more on experience-based intuition. This aligns with dual-process theorists' opinion that "expert knowledge" can lead to more intuition than deliberate reasoning in planning [56]. Thus, an agent may require a more powerful intuitive thinking system for affordance-based interaction planning tasks.

*6.3.2 Error Correction Phase.* We collected participants' subjective ratings of the original, improved, and transferred plans on a similar task regarding five metrics in Section 5.1.3. We conducted a within-subject non-parametric Friedman test followed by post hoc Wilcoxon signed-rank tests and presented the results in Fig. 15.

Overall, we found significant differences between the three types of plans across all metrics, except for respecting owner-defined usage-rights, as no plan violating this was selected. In pairwise comparison, the improved plans ($4.55 \pm 0.10$) scored significantly higher in overall acceptability than the original plans ($3.04 \pm 0.23$), with significant improvements in physical validity (from $3.55 \pm 0.32$ to $4.70 \pm 0.12$) and social appropriateness (from $3.44 \pm 0.24$ to $4.59 \pm 0.16$). The great standard deviation of the original plans' ratings reflected participants' diverse opinions when judging plans. In addition, the transferred plans to similar tasks received higher ratings ($4.94 \pm 0.23$ in overall acceptability) than the improved plans in all metrics though not significant. They significantly outperformed the original plans in planning success, physical validity, and social appropriateness (Fig. 15). This demonstrates that the *ACKnowledge* agent effectively learned from corrections and transferred this knowledge to similar tasks.

Our thematic analysis revealed that all participants were **satisfied with negotiating and making necessary rectifications with the agent, appreciating its responsiveness to their demands**. It is interesting to note that, although the reasons why the original plans were rated poorly varied among individuals, ranging from social inappropriateness to questionable physical validity, the participants generally agreed that **their concerns were addressed in the improved and transferred plans**. Participants also provided insightful advice on further enhancement. They suggested that **agents should self-learn from their own mistakes** (U3, U5, U8). As U8 stressed, *"It is good for ACKnowledge to gain knowledge from its error and the correction automatically."*

This expectation was also reflected in the participants' choice of error correction methods. We compared the usage of affordance correction, plan-with-mannual-rule correction, and plan-with-automatic-rule as mentioned in 4.2.2. Only two participants (U1, U4) attempted affordance correction. In contrast, others preferred making plan selection corrections. After reviewing alternative plans provided by the agent, most participants favored automatic rule generation for plan corrections, with only two choosing manual rule insertion.

*6.3.3 Personalization Phase.* Through behavior analysis, we first examined the types of personalization participants engaged in during this phase. All adjusted the usage-rights of objects, but only two (U2, U5) modified the affordance usages. The thematic analysis further revealed that our participants regarded **the adjustable features (usage-rights and affordance usages) provided as appropriate with acceptable interfaces**. Those who did not modify the affordance functions considered this feature to be *"dynamic and task-sensitive"* (U1) and could be *"changed later based on actual situations"* (U1). This aligns with our formative study findings that participants prefer agents to adapt to human habits gradually in real practices. The participants also confirmed that ***ACKnowledge*'s default perception of these features aligned with commonsense, though incomplete** (U1, U2, U3, U5, U9). They suggested that the default settings include diverse criteria like position (U5), hygiene requirements (U5), and object lifespan (U6). Additionally, some participants advocated more dynamic personalization options that adapt as tasks evolve, reflecting changing personal preferences (U1, U3).

Furthermore, the new interaction plans in the personalized home setting were rated highly acceptable overall ($4.67 \pm 0.17$). In qualitative feedback, all participants thought **the new plans were updated correctly to fit the new household settings**.

## 7 Discussion

We summarize key findings from our studies as below:

- Formative Study: Proposed DRs and instantiated computable factors in terms of physical, intrapersonal, and interpersonal contexts
- Study 2 - Simulations: *ACKnowledge* achieved a higher task execution disambiguity rate and significantly improved overall acceptability, particularly in usage-rights respectfulness and social appropriateness. Comparisons with baselines in tasks with/without environment updates highlighted the importance of context awareness and cognitive architecture for generating acceptable plans.

- Study 3 - Real-world settings: The planning process was evaluated by users as understandable, adaptable, and usable. Their behaviors and feedback indicated a preference for direct, user-friendly interfaces that require less manual input and offer greater personalizabilities.

Based on the findings, this section first discusses how our proposed *ACKnowledge* enhanced human-compatible interactions. Inspired by *ACKnowledge*'s performance in our studies, we identified future directions for more comprehensive and user-friendly frameworks aiming at long-horizon interaction planning. Lastly, we present the limitations and future works of this paper.

## 7.1 Towards Human-Compatible Agent-Environment Interaction

This paper conceptualizes human-compatible agent-environment interaction planning as a combination of acceptable planning results and an understandable planning process. Below, we reflect on the importance of these two elements.

*7.1.1 Human Compatibility Ensured by Acceptable Outcome.* In the results of our simulation study (Section 5.3), overall acceptability ($3.90 \pm 0.07$) of *ACKnowledge* is lower than plan success ($4.13 \pm 0.09$), and physical validity ($4.22 \pm 0.10$), with significant gaps found ($p = .019, p = .006$). Therefore, physically sound plans may not necessarily receive the same level of psychological acceptance. Results also showed that enhancing human-factors-related metrics like usage-rights respectfulness and social appropriateness contributed to more acceptable interaction planning. This aligns with previous research [28] emphasizing the need for both physical safety and psychological comfort in human-agent cooperation.

Humans' emphasis on psychological acceptance may stem from the enactive account of cognition, where they perceive environments based on what they can *do* [84]. Modern theories describe this process of unfolding possibilities for action as situated imagination [110]. When evaluating interaction feasibility with situated imagination, humans assess physical validity from a third-person view while considering psychological acceptance to ensure appropriateness from a first-person perspective [77]. Embodied with commonsense knowledge of affordance and awareness of context factors aligned with humans, *ACKnowledge* enhanced such psychological acceptance, thus improving human compatibility.

*7.1.2 Human Compatibility Facilitated by Transparent Planning Process.* While some researchers argue that the planning process is irrelevant to system users once results are satisfactory [75], others emphasize the importance of explainability and personalization through the interactions in the planning process. Thellman et al. [109] highlighted the importance of robots' explainability in social human-robot interactions, particularly regarding robots' beliefs and perceptions of environments. Lee et al. [64] concluded the benefits of personalization in terms of improving humans' satisfaction, rapport, cooperation, and engagement with the robots. Extending on these works and supported by the positive feedback from humans interacting with *ACKnowledge*, which enhances explainability and personalization through its "brain" module that computationally models human cognition 4.1.1, we postulate that transparency in planning improves human compatibility, particularly by fostering these two qualities.

There is no single optimal solution for all users in many scenarios. In the error correction phase of our case study, we observed that some participants accepted a proposed solution while others rejected it. For example, when the agent suggested using the owner's beige towel to clean broken glass, two participants agreed, seeing no restrictions implied by owner-exclusiveness, while seven declined due to hygiene concerns. This highlights the need for users to define and communicate their own optimal solutions. Black-box AI models (e.g., [46]) often rely on iterative training with sufficient input data covering different situations. In contrast, a transparent system like *ACKnowledge* incorporates user preferences from the start to prevent misalignment in planning, which is more data-efficient and communication-efficient [58, 78, 105, 115]. For one thing, a process aligned with human cognition enables users to identify errors in *ACKnowledge* and make direct corrections with a single-shot reference to the specific step(s) where errors occur (see Section 4.3.2). For another, such a system also allows users to access intermediate results, enabling them to select preferred alternatives without waiting for new plans to be fully generated, which is more cost-efficient. All these advantages boost personalization efficiency and increase users' trust in the agents. Several participants were surprised by *ACKnowledge*'s ability to provide alternative plans and transfer the single-shot corrections to new tasks. As U8 noted, *"As long as it seems to understand what I meant, I can trust it more,"* a sentiment echoed by four others.

## 7.2 Directions for Future Development of *ACKnowledge*-like Intelligent Agents

*7.2.1 Multimodal awareness of real-word contexts.* In the current implementation of *ACKnowledge*, we only considered factors that could be captured through visual detection or manual input (e.g., occupancy and usage-rights) and subsequently included them in the computational reasoning. While *ACKnowledge* outperformed baselines in affordance planning using the provided contextual information, we recognize the need to leverage broader contextual factors to enhance practical feasibility in robot deployments. Specifically, incorporating factors such as the physical properties of objects (e.g., mass, material) and environmental configurations (e.g., object positions, navigable pathways) can ensure that selected objects are accessible to the agents from their current location and practically feasible for the intended tasks. Feedback from participants also highlighted this need, for example, *"I may imply some usage-rights preference by the position of objects"* (U4) and *"Can this agent know the weight of the objects?"* (U1). One possible approach to achieving this is improving the *"vision"* module with multimodal contextual sensing and understanding [93, 122, 123]. Advanced sensing techniques [87, 99] that can capture richer contextual information to update KG structure and node properties [38] include but are not limited to capacitive proximity sensor [12], ultrasonic human activity sensor [40], piezoresistive pressure sensor [16]) and multimodal foundation models [66]. In addition, using 3D object sensors and scene reconstruction algorithms, we can model the environment as a 3D scene graph (e.g., [81, 94, 101]). This will enable path and motion planning with spatial information rather

than merely validating the feasibility of affordance choice at the destination.

*7.2.2 Temporal-Spatial Long-horizon Interaction Planning.* *ACKnowledge* showed the satisfactory capability of planning object interactions in independent tasks (Section 5.3) with the potential to extend to planning long-horizon interactions. Human interactions are relational and influenced by both time and space [51]. From a temporal viewpoint, future agents should be able to plan a series of interactions for long-horizon tasks in the evolving "landscape" [91, 128]. For example, when "arranging a dinner party", they should consider how earlier actions might alter the environment (e.g., occupancy, location, etc.) for subsequent activities. The agents have to balance the trade-offs to ensure plausible outcomes in each step and the whole interaction sequence while maintaining transparency and adaptability throughout the process [27, 100]. Successful long-horizon planning will also accumulate data on users' preferences and habits under multiple constraints and objectives, helping agents self-train their planning models to better adapt to human characteristics. From a space perspective, future agents can extend *ACKnowledge*'s ability to plan interactions beyond households (e.g., workplaces, restaurants, etc.) and across locations, allowing for interaction without geographical constraints. For instance, when asked to make coffee for the host, an agent should find, clean, and use the host's usual cup instead of fetching a new one each time. This necessitates the agent tracking object locations and understanding individual and social preferences for object use and arrangement [46, 67]. With such capabilities, future agents could offer users a more connected and convenient experience across environments.

*7.2.3 Interaction potential between intelligent agents and users.*
**User-friendly interaction with the agent**. In the case study, we offered the users a personalization web app GUI with simple text or button input and a keyword-based error correction CUI. While participants appreciated the functionality of *ACKnowledge*'s interfaces, they suggested improvements for user-friendliness. They would like to explain their preferences and rectify undesirable actions by show-and-tell as they walk the agent through the environment. Existing research also indicates that simple demonstrations in daily activities [98] and sparse feedback [89] enhance usability. *"It is better if I can tell them my demands just in a few sentences"* (U1, U5, U9). In this way, users can directly showcase their demands with less mental load, leaving the work of interpreting humans' needs and intentions to the agents.

**Multi-user interaction with the agent**. In both the design and the evaluation of *ACKnowledge*, we mainly focused on the subjective ratings of the homeowner and received satisfactory responses (Section 5.3). However, shared environments include individuals with diverse identities and preferences. Users may prefer different solutions for the same planning task based on their emphasis on various criteria. As shown in the results from Section 6.3.2, users often interpret the importance of each metric differently when evaluating a plan. Additionally, they may favor different interaction styles (verbal v.s. physical [108], interaction distance [96]) with the agent. Intelligent agents should accommodate all users and minimize conflicts over preferences.

They should also promote positive interactions between humans sharing the environment [42, 61], which has long been recognized in Human-Agent Teaming practices [83], such as reminding users about others' disliked behaviors. As participant P11 suggested, *"I would like the agent to plan some activities that can improve my relationship with my guest."* Thus, extending *ACKnowledge* to meet the varied needs of multiple users is crucial and warrants further research.

## 7.3 Limitation and Future Works

This work has several limitations. First, we developed the foundational affordance commonsense knowledge graph (KG) that supports the perception and reasoning of *ACKnowledge* solely using the *ObjectUse* tuples from the ATOMIC2020 dataset [52]. This single-source KG may be incomplete (e.g., missing object/affordance nodes and other types of relations) and biased by the available crowdsourcing data. Additionally, the KG-based reasoning approach in our work relies primarily on retrieval. Second, we conducted an imaginary formative study with a simulation evaluation and a case study using an AR agent instead of a real robot. In the scope of this paper, we explored how a robot's mind should function in interaction planning. We did not consider actual robot motion capabilities and factors beyond household tasks. Third, our studies instructed the participants to assume that all the information floating through *ACKnowledge* was secure and privacy-preserving without investigating their perception of risks and potential protective measures. Additionally, the generalizability of our results is constrained by the small sample sizes and the narrow age range of participants.

To overcome these limitations, in the future, we will incorporate data from more diverse sources (e.g., Wikidata [111], Visual Genome [63]) and integrate subjective relations, such as human intentions, to construct a KG in line with human perception and cognitive structure. We plan to use recent LLMs with enhanced commonsense reasoning to update our KG with emerging objects and affordance, such as by suggesting links and weights based on similarity with existing nodes and edges. Besides, LLMs can prompt action plans according to inferred user intentions without explicit instruction. For instance, agents may suggest "setting the table" when a user mentions hunger.

Furthermore, we plan to deploy *ACKnowledge* on real robots, with privacy protection and security of personal information taken into consideration, to validate our findings on user expectations of agents. Participants will instruct a robot to carry out affordance-based tasks and evaluate its plans and executions with metrics employed in this paper. The results will verify the effectiveness of various system components built upon the formative findings. We will also conduct interviews to assess whether the robot meets expectations and gather insights for future improvement. Finally, we aim to expand our participant groups with more task scenarios in future studies to enhance the generalizability of our results.

## 8 Conclusion

In this work, we conducted a formative study identifying the physical, intrapersonal, and interpersonal contexts that count to household human-agent interaction. Extending the existing emphasis

on explainability and personalizability in social human-robot interaction research, we then proposed *ACKnowledge*, a computational framework integrating a dynamic knowledge graph, a large language model, and a vision language model as the *"brain"* and *"vision"* to articulate the dual-process reasoning with retrieval augmented generation for affordance-based interaction planning in dynamic human environments. *ACKnowledge* demonstrated acceptable planning results through an understandable process through evaluations. In real-world simulation tasks, it attained high execution disambiguity and overall acceptability, significantly improving respect for usage rights and social appropriateness compared to baseline models. Feedback from the case study highlighted *ACKnowledge*'s ability to negotiate and personalize within an understandable planning framework. This work offers valuable insights into human-compatible agent-environment interactions, merging cognitive theories with human perceptions and inspiring future advancements in human-agent coexistence.

## Acknowledgments

This work is supported by the Research Grants Council of the Hong Kong Special Administrative Region under the General Research Fund (GRF) with Grant No. 16207923.

Thanks to the UROP students Yeung Kong, Lam, and Wing Ip, Lo, for contributing to the project during their summer research internship. Thank you to everyone who lifted me through this endless summer; your support helped me soar and brightened my days. To all the voices, chords, and melodies that have either haunted or embraced, killed or saved me, you are indispensable parts of my life. Hard times come again no more.

## References

[1] 1999. Case studies of applying Gibson's ecological approach to mobile robots. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 29, 1 (1999), 105–111.

[2] Henk Aarts and Ap Dijksterhuis. 2003. The silence of the library: environment, situational norm, and social behavior. *Journal of personality and social psychology* 84, 1 (2003), 18.

[3] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. arXiv:2204.01691 [cs.RO]

[4] Rachid Alami, Aurélie Clodic, Vincent Montreuil, Emrah Akin Sisbot, and Raja Chatila. 2006. Toward Human-Aware Robot Task Planning.. In *AAAI spring symposium: to boldly go where no human-robot team has gone before*. 39–46.

[5] Mohammad Rafayet Ali, Seyedeh Zahra Razavi, Raina Langevin, Abdullah Al Mamun, Benjamin Kane, Reza Rawassizadeh, Lenhart K Schubert, and Ehsan Hoque. 2020. A virtual conversational agent for teens with autism spectrum disorder: Experimental results and design lessons. In *Proceedings of the 20th ACM international conference on intelligent virtual agents*. 1–8.

[6] Ettore Ambrosini, Claudia Scorolli, Anna M Borghi, and Marcello Costantini. 2012. Which body for embodied cognition? Affordance and language within actual and perceived reaching space. *Consciousness and cognition* 21, 3 (2012), 1551–1557.

[7] Kamiar Aminian and Bijan Najafi. 2004. Capturing human motion using body-fixed sensors: outdoor measurement and clinical applications. *Computer animation and virtual worlds* 15, 2 (2004), 79–94.

[8] R Arora, S Singh, K Swaminathan, A Datta, S Banerjee, B Bhowmick, KM Jatavallabhula, and M Sridharan. [n. d.]. Anticipate & Act: Integrating LLMs and Classical Planning for Efficient Task Execution in Household Environments. In the IEEE. ([n. d.]).

[9] Susy Budi Astuti, Purwanita Setijanti, and Ispurwono Soemarno. 2017. Personalization of space in private and public setting within vertical housing as sustainable living. *DIMENSI (Journal of Architecture and Built Environment)* 44, 1 (2017), 37–44.

[10] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. 2023. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13778–13790.

[11] Martin Benfeghoul, Umais Zahid, Qinghai Guo, and Zafeirios Fountas. 2024. When in Doubt, Think Slow: Iterative Reasoning with Latent Imagination. *arXiv preprint arXiv:2402.15283* (2024).

[12] Miodrag Bolic, Majed Rostamian, and Petar M Djuric. 2015. Proximity detection with RFID: A step toward the internet of things. *IEEE Pervasive Computing* 14, 2 (2015), 70–76.

[13] Anna M Borghi. 2021. Affordances, context and sociality. *Synthese* 199, 5 (2021), 12485–12515.

[14] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[15] Sandra Brunia and Anca Hartjes-Gosselink. 2009. Personalization in non-territorial offices: a study of a human need. *Journal of Corporate Real Estate* 11, 3 (2009), 169–182.

[16] Wei Cao, Yan Luo, Yiming Dai, Xin Wang, Kaili Wu, Huijuan Lin, Kun Rui, and Jixin Zhu. 2023. Piezoresistive pressure sensor based on a conductive 3D sponge network for motion sensing and human–machine interface. *ACS Applied Materials & Interfaces* 15, 2 (2023), 3131–3140.

[17] Filippo Cavallo, Raffaele Limosani, Alessandro Manzi, Manuele Bonaccorsi, Raffaele Esposito, Maurizio Di Rocco, Federico Pecora, Giancarlo Teti, Alessandro Saffiotti, and Paolo Dario. 2014. Development of a socially believable multi-robot solution from town to home. *Cognitive Computation* 6 (2014), 954–967.

[18] Janis Chadsey and Steve Beyer. 2001. Social relationships in the workplace. *Mental retardation and developmental disabilities research reviews* 7, 2 (2001), 128–133.

[19] Georgia Chalvatzaki, Ali Younes, Daljeet Nandha, An Thai Le, Leonardo FR Ribeiro, and Iryna Gurevych. 2023. Learning to reason over scene graphs: a case study of finetuning GPT-2 into a robot language model for grounded task planning. *Frontiers in Robotics and AI* 10 (2023), 1221739.

[20] Anthony Chemero. 2018. An outline of a theory of affordances. In *How Shall Affordances Be Refined?* Routledge, 181–195.

[21] Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, Keerthana Gopalakrishnan, Michael S. Ryoo, Austin Stone, and Daniel Kappler. 2023. Open-vocabulary Queryable Scene Representations for Real World Planning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. 11509–11522. https://doi.org/10.1109/ICRA48891.2023.10161534

[22] Yi Fei Cheng, Christoph Gebhardt, and Christian Holz. 2023. Interactionadapt: Interaction-driven workspace adaptation for situated virtual reality environments. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[23] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. 2018. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 975–983.

[24] Mina Cikara, Joel E Martinez, and Neil A Lewis Jr. 2022. Moving beyond social categories by incorporating context in social psychological theory. *Nature Reviews Psychology* 1, 9 (2022), 537–549.

[25] Marcello Cirillo, Lars Karlsson, and Alessandro Saffiotti. 2010. Human-aware task planning: An application to mobile robots. *ACM Transactions on Intelligent Systems and Technology (TIST)* 1, 2 (2010), 1–26.

[26] Alan Costall and Ann Richards. 2013. Canonical affordances: The psychology of everyday things. *The Oxford handbook of the archaeology of the contemporary world* (2013), 82–93.

[27] Murtaza Dalal, Tarun Chiruvolu, Devendra Chaplot, and Ruslan Salakhutdinov. 2024. Plan-seq-learn: Language model guided rl for solving long horizon robotics tasks. *arXiv preprint arXiv:2405.01534* (2024).

[28] Dayle David, Pierre Thérouanne, and Isabelle Milhabet. 2022. The acceptability of social robots: A scoping review of the recent literature. *Computers in Human Behavior* 137 (2022), 107419.

[29] Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 2021. 3D AffordanceNet: A Benchmark for Visual Object Affordance Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1778–1787.

[30] Thanh-Toan Do, Anh Nguyen, and Ian Reid. 2018. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 5882–5889.

[31] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. PaLM-E: An

Embodied Multimodal Language Model. arXiv:2303.03378 [cs.LG] https://arxiv.org/abs/2303.03378

[32] Ruofei Du, Alex Olwal, Mathieu Le Goc, Shengzhi Wu, Danhang Tang, Yinda Zhang, Jun Zhang, David Joseph Tan, Federico Tombari, and David Kim. 2022. Opportunistic Interfaces for Augmented Reality: Transforming Everyday Objects into Tangible 6DoF Interfaces Using Ad hoc UI. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 183, 4 pages. https://doi.org/10.1145/3491101.3519911

[33] Wanyu Du, Zae Myung Kim, Vipul Raheja, Dhruv Kumar, and Dongyeop Kang. 2022. Read, revise, repeat: A system demonstration for human-in-the-loop iterative text revision. *arXiv preprint arXiv:2204.03685* (2022).

[34] Andrew P Duchon, Leslie Pack Kaelbling, and William H Warren. 1998. Ecological robotics. *Adaptive Behavior* 6, 3-4 (1998), 473–507.

[35] Florian Dufresne, Charlotte Dubosc, Geoffrey Gorisse, and Olivier Christmann. 2024. Understanding the Impact of Coherence between Virtual Representations and Possible Interactions on Embodiment in VR: an Affordance Perspective. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[36] Cathy Mengying Fang, Ryo Suzuki, and Daniel Leithinger. 2023. VR Haptics at Home: Repurposing Everyday Objects and Environment for Casual and On-Demand VR Haptic Experiences. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 312, 7 pages. https://doi.org/10.1145/3544549.3585871

[37] Sean Follmer, Daniel Leithinger, Alex Olwal, Akimitsu Hogge, and Hiroshi Ishii. 2013. inFORM: dynamic physical affordances and constraints through shape and object actuation.. In *Uist*, Vol. 13. Citeseer, 2501–988.

[38] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. 2024. Physically Grounded Vision-Language Models for Robotic Manipulation. arXiv:2309.02561 [cs.RO] https://arxiv.org/abs/2309.02561

[39] Wanting Gao, Xinyi Gao, and Yin Tang. 2023. Multi-Turn Dialogue Agent as Sales' Assistant in Telemarketing. In *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–9.

[40] Arindam Ghosh, Amartya Chakraborty, Dhruv Chakraborty, Mousumi Saha, and Sujoy Saha. 2023. UltraSense: A non-intrusive approach for human activity identification using heterogeneous ultrasonic sensor grid for smart home environment. *Journal of Ambient Intelligence and Humanized Computing* (2023), 1–22.

[41] James J Gibson. 1977. The theory of affordances. *Hilldale, USA* 1, 2 (1977), 67–82.

[42] Sarah Gillet, Marynel Vázquez, Sean Andrist, Iolanda Leite, and Sarah Sebo. 2024. Interaction-Shaping Robotics: Robots That Influence Interactions between Other Agents. *J. Hum.-Robot Interact.* 13, 1, Article 12 (March 2024), 23 pages. https://doi.org/10.1145/3643803

[43] Weilun Gong, Stephanie Santosa, Tovi Grossman, Michael Glueck, Daniel Clarke, and Frances Lai. 2023. Affordance-Based and User-Defined Gestures for Spatial Tangible Interaction. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 1500–1514.

[44] Moritz A Graule and Volkan Isler. 2024. Gg-llm: Geometrically grounding large language models for zero-shot human activity forecasting in human-aware task planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 568–574.

[45] Omer Gvirsman, Yaacov Koren, Tal Norman, and Goren Gordon. 2020. Patricc: A platform for triadic interaction with changeable characters. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-robot Interaction*. 399–407.

[46] Dongge Han, Trevor McInroe, Adam Jelley, Stefano V Albrecht, Peter Bell, and Amos Storkey. 2024. LLM-Personalize: Aligning LLM Planners with Human Preferences via Reinforced Self-Training for Housekeeping Robots. *arXiv preprint arXiv:2404.14285* (2024).

[47] Rishi Hazra, Pedro Zuidberg Dos Martires, and Luc De Raedt. 2024. Saycanpay: Heuristic planning with large language models using learnable domain knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 20123–20133.

[48] Marcel Heerink, Ben Kröse, Vanessa Evers, and Bob Wielinga. 2010. Assessing acceptance of assistive social agent technology by older adults: the almere model.

[49] Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. 2023. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv:2311.17842* (2023).

[50] Wenlong Huang, Fei Xia, Dhruv Shah, Danny Driess, Andy Zeng, Yao Lu, Pete Florence, Igor Mordatch, Sergey Levine, Karol Hausman, et al. 2024. Grounded decoding: Guiding text generation with grounded models for embodied agents. *Advances in Neural Information Processing Systems* 36 (2024).

[51] Edwin Hutchins. 2020. The distributed cognition perspective on human interaction. In *Roots of human sociality*. Routledge, 375–398.

[52] Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 6384–6392.

[53] Filip Ilievski, Pedro Szekely, and Bin Zhang. 2021. Cskg: The commonsense knowledge graph. In *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18*. Springer, 680–696.

[54] Rahul Jain, Jingyu Shi, Runlin Duan, Zhengzhe Zhu, Xun Qian, and Karthik Ramani. 2023. Ubi-TOUCH: Ubiquitous Tangible Object Utilization through Consistent Hand-object interaction in Augmented Reality. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 12, 18 pages. https://doi.org/10.1145/3586183.3606793

[55] Haiyan Jiang, Dongdong Weng, Xiaonuo Dongye, Nan Zhang, and Luo Le. 2023. A commonsense knowledge-based object retrieval approach for Virtual reality. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 795–796.

[56] Daniel Kahneman. 2011. Thinking Fast and Slow. *Farrar, Strauss and Giroux* (2011).

[57] Victor Kaptelinin and Bonnie Nardi. 2012. Affordances in HCI: toward a mediated action perspective. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 967–976.

[58] Laurent Karsenty and Valérie Botherel. 2005. Transparency strategies to help users handle system errors. *Speech Communication* 45, 3 (2005), 305–324.

[59] Junghyun Kim, Gi-Cheon Kang, Jaein Kim, Seoyun Yang, Minjoon Jung, and Byoung-Tak Zhang. 2023. PGA: Personalizing Grasping Agents with Single Human-Robot Interaction. *arXiv preprint arXiv:2310.12547* (2023).

[60] Kobe Knowles, Michael Witbrock, Gillian Dobbie, and Vithya Yogarajan. 2023. A Proposal for a Language Model Based Cognitive Architecture. In *Proceedings of the AAAI Symposium Series*, Vol. 2. 295–301.

[61] Kanae Kochigami, Kei Okada, and Masayuki Inaba. 2021. Pilot Study on Robot's Open Diary to Deepen Friendships with a Child and Promote Communication between a Child and People. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 104–108.

[62] Uwe Köckemann, Federico Pecora, and Lars Karlsson. 2014. Grandpa hates robots-interaction constraints for planning in inhabited environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 28.

[63] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123 (2017), 32–73.

[64] Min Kyung Lee, Jodi Forlizzi, Sara Kiesler, Paul Rybski, John Antanitis, and Sarun Savetsila. 2012. Personalization in HRI: A longitudinal field experiment. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. 319–326.

[65] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023. MIMIC-IT: Multi-Modal In-Context Instruction Tuning. arXiv:2306.05425 [cs.CV] https://arxiv.org/abs/2306.05425

[66] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. 2024. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision* 16, 1-2 (2024), 1–214.

[67] Wenhao Li, Zhiyuan Yu, Qijin She, Zhinan Yu, Yuqing Lan, Chenyang Zhu, Ruizhen Hu, and Kai Xu. 2024. LLM-enhanced Scene Graph Learning for Household Rearrangement. *arXiv preprint arXiv:2408.12093* (2024).

[68] Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. 2019. Putting humans in a scene: Learning affordance in 3d indoor environments. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12368–12376.

[69] Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. 2023. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. *arXiv preprint arXiv:2305.13269* (2023).

[70] Yi-Chi Liao, Kashyap Todi, Aditya Acharya, Antti Keurulainen, Andrew Howes, and Antti Oulasvirta. 2022. Rediscovering affordance: A reinforcement learning perspective. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.

[71] Bill Yuchen Lin, Yicheng Fu, Karina Yang, Faeze Brahman, Shiyu Huang, Chandra Bhagavatula, Prithviraj Ammanabrolu, Yejin Choi, and Xiang Ren. 2024. Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks. *Advances in Neural Information Processing Systems* 36 (2024).

[72] Chuan-en Lin, Ta Ying Cheng, and Xiaojuan Ma. 2020. Architect: Building interactive virtual experiences from physical affordances by bringing human-in-the-loop. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[73] Yuchen Liu, Luigi Palmieri, Sebastian Koch, Ilche Georgievski, and Marco Aiello. 2024. Towards Human Awareness in Robot Task Planning with Large Language

Models. *arXiv preprint arXiv:2404.11267* (2024).

[74] Christopher Lörken and Joachim Hertzberg. 2008. Grounding planning operators by affordances. In *International Conference on Cognitive Systems (CogSys)*. Citeseer, 79–84.

[75] Octavio Loyola-Gonzalez. 2019. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE access* 7 (2019), 154096–154113.

[76] Liangsheng Lu, Wei Zhai, Hongchen Luo, Yu Kang, and Yang Cao. 2022. Phrase-based affordance detection via cyclic bilateral interaction. *IEEE Transactions on Artificial Intelligence* 4, 5 (2022), 1186–1198.

[77] Tom McClelland. 2019. Representing our options: The perception of affordances for bodily and mental action. *Journal of Consciousness Studies* 26, 3-4 (2019), 155–180.

[78] Joseph E Mercado, Michael A Rupp, Jessie YC Chen, Daniel Barber, Katelyn Procci, and Michael Barnes. 2015. Effects of agent transparency on multi-robot management effectiveness. *Aberdeen Proving Ground (MD): Army Research Laboratory (US)* (2015).

[79] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. 2019. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 909–918.

[80] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 2024. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems* 36 (2024).

[81] Francisco Munguia-Galeano, Satheeshkumar Veeramani, Juan David Hernández, Qingmeng Wen, and Ze Ji. 2023. Affordance-based human–robot interaction with reinforcement learning. *IEEE Access* 11 (2023), 31282–31292.

[82] Yasuto Nakanishi. 2022. Furnituroid: Shape-Changing Mobile Furniture Robot for Multiple and Dynamic Affordances. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 937–939.

[83] Manisha Natarajan, Esmaeil Seraj, Batuhan Altundas, Rohan Paleja, Sean Ye, Letian Chen, Reed Jensen, Kimberlee Chestnut Chang, and Matthew Gombolay. 2023. Human-robot teaming: grand challenges. *Current Robotics Reports* 4, 3 (2023), 81–100.

[84] Albert Newen, Leon De Bruin, and Shaun Gallagher. 2018. *The Oxford handbook of 4E cognition*. Oxford University Press.

[85] Benjamin A Newman, Pranay Gupta, Yonatan Bisk, Kris Kitani, Henny Admoni, and Chris Paxton. 2024. Leveraging Vision and Language Models for Zero-Shot, Personalization of Household Multi-Object Rearrangement Tasks. (2024).

[86] Donald A Norman. 1999. Affordance, conventions, and design. *interactions* 6, 3 (1999), 38–43.

[87] Vasilios A Orfanos, Stavros D Kaminaris, Panagiotis Papageorgas, Dimitrios Piromalis, and Dionisis Kandris. 2023. A comprehensive review of IoT networking technologies for smart home automation applications. *Journal of Sensor and Actuator Networks* 12, 2 (2023), 30.

[88] Lucas Paletta, Gerald Fritz, Florian Kintzler, Jorg Irran, and Georg Dorffner. 2007. Learning to perceive affordances in a framework of developmental embodied cognition. In *2007 IEEE 6th International Conference on Development and Learning*. IEEE, 110–115.

[89] Maithili Patel and Sonia Chernova. 2024. Robot Behavior Personalization from Sparse User Feedback. *arXiv preprint arXiv:2410.19219* (2024).

[90] Siyou Pei, Alexander Chen, Jaewook Lee, and Yang Zhang. 2022. Hand interfaces: Using hands to imitate objects in ar/vr for expressive interactions. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–16.

[91] Sören Pirk, Karol Hausman, Alexander Toshev, and Mohi Khansari. 2020. Modeling Long-horizon Tasks as Sequential Interaction Landscapes. arXiv:2006.04843 [cs.RO] https://arxiv.org/abs/2006.04843

[92] Jielin Qiu, Andrea Madotto, Zhaojiang Lin, Paul A Crook, Yifan Ethan Xu, Xin Luna Dong, Christos Faloutsos, Lei Li, Babak Damavandi, and Seungwhan Moon. 2024. Snapntell: Enhancing entity-centric visual question answering with retrieval augmented multimodal llm. *arXiv preprint arXiv:2403.04735* (2024).

[93] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. 2018. Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 4 (2018), 1–27.

[94] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian D. Reid, and Niko Sünderhauf. 2023. SayPlan: Grounding Large Language Models using 3D Scene Graphs for Scalable Task Planning. *CoRR* abs/2307.06135 (2023). https://doi.org/10.48550/arXiv.2307.06135

[95] Erik Rietveld, Damiaan Denys, and Maarten Van Westen. 2018. Ecological-enactive cognition as engaging with a field of relevant affordances. *The Oxford handbook of 4E cognition* 41 (2018), 70.

[96] Lionel P Robert Jr, Rasha Alahmad, Connor Esterwood, Sangmi Kim, Sangseok You, Qiaoning Zhang, et al. 2020. A review of personality in human–robot interactions. *Foundations and Trends® in Information Systems* 4, 2 (2020), 107–212.

[97] Erol Şahin, Maya Cakmak, Mehmet R Doğar, Emre Uğur, and Göktürk Üçoluk. 2007. To afford or not to afford: A new formalization of affordances toward affordance-based robot control. *Adaptive Behavior* 15, 4 (2007), 447–472.

[98] Joe Saunders, Dag Sverre Syrdal, Kheng Lee Koay, Nathan Burke, and Kerstin Dautenhahn. 2015. "teach me–show me"—end-user personalization of a smart home and companion robot. *IEEE Transactions on Human-Machine Systems* 46, 1 (2015), 27–40.

[99] Deepti Sehrawat and Nasib Singh Gill. 2019. Smart sensors: Analysis of different types of IoT sensors. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 523–528.

[100] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. 2024. Robovqa: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 645–652.

[101] Zachary Seymour, Niluthpol Chowdhury Mithun, Han-Pang Chiu, Supun Samarasekera, and Rakesh Kumar. 2022. GraphMapper: Efficient Visual Navigation by Scene Graph Generation. In *2022 26th International Conference on Pattern Recognition (ICPR)*. 4146–4153. https://doi.org/10.1109/ICPR56361.2022.9956224

[102] Bruce Sherin. 2006. Common sense clarified: The role of intuitive knowledge in physics problem solving. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching* 43, 6 (2006), 535–555.

[103] John Shotter. 1983. Duality of structure" and "intentionality" in an ecological psychology. *Journal for the Theory of Social Behaviour* 13, 1 (1983), 19–44.

[104] Austin Stone, Ted Xiao, Yao Lu, Keerthana Gopalakrishnan, Kuang-Huei Lee, Quan Vuong, Paul Wohlhart, Sean Kirmani, Brianna Zitkovich, Fei Xia, et al. 2023. Open-world object manipulation using pre-trained vision-language models. *arXiv preprint arXiv:2303.00905* (2023).

[105] Kimberly Stowers, Nicholas Kasdaglis, Michael A Rupp, Olivia B Newton, Jessie YC Chen, and Michael J Barnes. 2020. The IMPACT of agent transparency on human performance. *IEEE Transactions on Human-Machine Systems* 50, 3 (2020), 245–253.

[106] Theodore R Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. 2023. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427* (2023).

[107] Ron Sun. 2024. Can A Cognitive Architecture Fundamentally Enhance LLMs? Or Vice Versa? *arXiv preprint arXiv:2401.10444* (2024).

[108] Dag Sverre Syrdal, Kheng Lee Koay, Michael L. Walters, and Kerstin Dautenhahn. 2007. A personalized robot companion? - The role of individual differences on spatial preferences in HRI scenarios. In *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication*. 1143–1148. https://doi.org/10.1109/ROMAN.2007.4415252

[109] Sam Thellman and Tom Ziemke. 2021. The perceptual belief problem: Why explainability is a tough challenge in social robotics. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 3 (2021), 1–15.

[110] Ludger Van Dijk and Erik Rietveld. 2020. Situated imagination. *Phenomenology and the Cognitive Sciences* (2020), 1–23.

[111] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[112] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2023. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214* (2023).

[113] Peter C Wason and J St BT Evans. 1974. Dual processes in reasoning? *Cognition* 3, 2 (1974), 141–154.

[114] Elke U Weber, Sara M Constantino, and Maja Schlüter. 2023. Embedding cognition: judgment and choice in an interdependent and dynamic world. *Current Directions in Psychological Science* 32, 4 (2023), 328–336.

[115] Julia L Wright, Jessie YC Chen, and Shan G Lakhmani. 2019. Agent transparency and reliability in human–robot interaction: The influence on user confidence and perceived reliability. *IEEE Transactions on Human-Machine Systems* 50, 3 (2019), 254–263.

[116] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. 2023. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots* 47, 8 (2023), 1087–1102.

[117] Shengzhi Wu, Daragh Byrne, and Molly Wright Steenson. 2020. " Megereality": Leveraging Physical Affordances for Multi-Device Gestural Interaction in Augmented Reality. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–4.

[118] Zhanxin Wu, Bo Ai, and David Hsu. 2023. Integrating Common Sense and Planning with Large Language Models for Room Tidying. In *RSS 2023 Workshop on Learning for Task and Motion Planning*.

[119] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864* (2023).

[120] Kosei Yamao, Daiju Kanaoka, Kosei Isomoto, Akinobu Mizutani, Yuichiro Tanaka, and Hakaru Tamukoh. 2024. Development of A SayCan-based Task Planning System Capable of Handling Abstract Nouns. In *Proceedings of International Conference on Artificial Life & Robotics (ICAROB2024)*. ALife Robotics, OS15–4.

[121] Yukang Yan, Chun Yu, Xiaojuan Ma, Xin Yi, Ke Sun, and Yuanchun Shi. 2018. Virtualgrasp: Leveraging experience of interacting with physical objects to facilitate digital object retrieval. In *Proceedings of the 2018 Chi conference on human factors in computing systems*. 1–13.

[122] Jing Yang, Nade Liang, Brandon J Pitts, Kwaku O Prakah-Asante, Reates Curry, Mike Blommer, Radhakrishnan Swaminathan, and Denny Yu. 2023. Multimodal sensing and computational intelligence for situation awareness classification in autonomous driving. *IEEE Transactions on Human-Machine Systems* 53, 2 (2023), 270–281.

[123] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. 2024. Contextual object detection with multimodal large language models. *International Journal of Computer Vision* (2024), 1–19.

[124] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. 2023. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075* (2023).

[125] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2024. MME-RealWorld: Could Your Multimodal LLM Challenge High-Resolution Real-World Scenarios that are Difficult for Humans? arXiv:2408.13257 [cs.CV] https://arxiv.org/abs/2408.13257

[126] Wangchunshu Zhou, Yuchen Eleanor Jiang, Long Li, Jialong Wu, Tiannan Wang, Shi Qiu, Jintian Zhang, Jing Chen, Ruipu Wu, Shuai Wang, et al. 2023. Agents: An open-source framework for autonomous language agents. *arXiv preprint arXiv:2309.07870* (2023).

[127] Yuqi Zhu, Shuofei Qiao, Yixin Ou, Shumin Deng, Ningyu Zhang, Shiwei Lyu, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024. Knowagent: Knowledge-augmented planning for llm-based agents. *arXiv preprint arXiv:2403.03101* (2024).

[128] Yifeng Zhu, Jonathan Tremblay, Stan Birchfield, and Yuke Zhu. 2021. Hierarchical Planning for Long-Horizon Manipulation with Geometric and Symbolic Scene Graphs. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. 6541–6548. https://doi.org/10.1109/ICRA48506.2021.9561548