
COOPERA: Continual Open-Ended Human-Robot Assistance

Chenyang Ma¹ Kai Lu¹
Ruta Desai[†] Xavier Puig[†] Andrew Markham^{1†} Niki Trigoni^{1†}

¹University of Oxford

Abstract

To understand and collaborate with humans, robots must account for individual human traits, habits, and activities over time. However, most robotic assistants lack these abilities, as they primarily focus on predefined tasks in structured environments and lack a human model to learn from. This work introduces **COOPERA**, a novel framework for COntinual, OPen-Ended human-Robot Assistance, where simulated humans, driven by psychological traits and long-term intentions, interact with robots in complex environments. By integrating continuous human feedback, our framework, for the first time, enables the study of long-term, open-ended human-robot collaboration (HRC) in different collaborative tasks across various time-scales. Within COOPERA, we introduce a benchmark and an approach to personalize the robot’s collaborative actions by learning human traits and context-dependent intents. Experiments validate the extent to which our simulated humans reflect realistic human behaviors and demonstrate the value of inferring and personalizing to human intents for open-ended and long-term HRC.

1 Introduction

A long-standing goal in robotics is to develop agents that can effectively assist humans in their daily lives by adapting to their preferences and habits. In order to do this, a robot agent must be able to not only learn to interact in environments with humans in a given moment, but also reason about the human across long periods of time, adapting its behavior to provide better assistance. For example, such an agent should be able to fetch a cup of coffee while also understanding that someone may prefer it cooler in the morning but stronger in the afternoon, heating up water accordingly.

Over recent years, several works have made significant advances in developing agents that can assist humans in household tasks [52, 50, 75, 65, 47], using simulation environments to study human-robot collaboration (HRC) in a safe and scalable manner. However, most of these works focus on episodic settings, where a robot is evaluated over a set of short collaboration scenarios with tasks specified in advance. These settings are very different from real-world scenarios, where humans have preferences and long-term goals that guide their behaviors, needing different types of assistance at different times.

To advance robot agents that can assist and adapt to humans, we propose **COOPERA**, a novel framework for COntinual, OPen-Ended human-Robot Assistance in complex household environments (Fig. 1). At its core, COOPERA features a human model with preferences that supports long-term interactions, a feedback mechanism, and benchmarks and metrics to evaluate if robots can assist humans in long-term tasks and reason about their preferences effectively.

To model realistic humans, we simulate humans using an LLM with detailed human traits and habits, retrieved environment information, and intention history, enabling behaviors that exhibit three characteristics. **1) Dynamic intention-driven:** Humans act based on intentions that vary over time

[†]Equal advising.

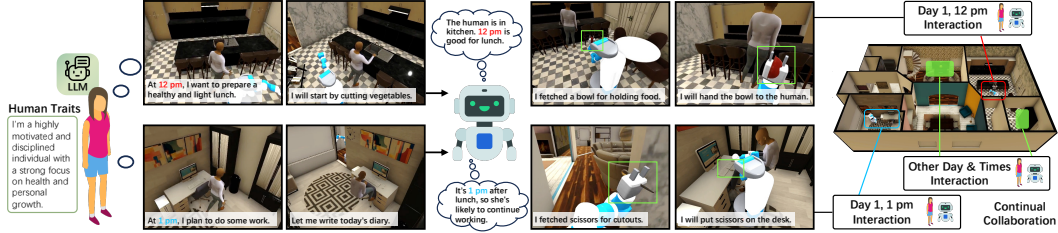


Figure 1: **Continual human-robot collaboration for open-ended tasks over multiple days.** Our framework **COOPERA** entails an approach to simulate traits-driven humans with long-term, whole-day behaviors within robot simulation platform, enabling the first study of long-term, open-ended human-robot collaboration. We also introduce a benchmark and a method for the robot to personalize collaboration in such continual, open-ended settings by learning human traits and context-dependent intents over time.

(e.g., setting the dinner table at 6 pm, then watching TV at 7 pm). **2) Open-ended and environment-conditioned:** Rather than following predefined tasks, humans generate spontaneous intentions based on the environment, available objects, and time of day. **3) Traits-driven:** Psychological traits and habits shape human behavior, resulting in diverse routines even within similar environments (e.g., one person starts their day reading, while another prefers cleaning).

As the human interacts in the environment, we need a way to provide feedback to the robot so it can improve over time. We structure our framework into two stages which happen on each day of interaction. At the beginning of the day, the robot observes and collaborates with the human, assisting in inferred tasks. At the day’s end, the human communicates with the robot and provides feedback to help improve the robot’s collaboration success rate for subsequent days.

COOPERA presents unique challenges that are often overlooked in existing HRC benchmarks. First, robot agents need to reason not only about the environment state but also a given human’s behavior for effective assistance. Rather than learning a behavior that assists all possible humans, they need to adapt to each person’s preferences and traits. Second, humans in our framework perform different tasks depending on the time of the day or the day of the week (e.g., they may only do exercise once). Thus, effective agents need to reason about time as a cue for how to best assist the human.

To explore this challenging framework, we provide a benchmark and propose a method that tracks human preference profiles and uses VLMs and classifiers to predict and score human goals based on time, observed behavior, and profile, suggesting actions to achieve these intentions. This enables the robot to learn and mimic human behavior by capturing the underlying correlations between human traits, temporal dependencies, and their corresponding intentions and tasks. We compare our method against several baselines, evaluating the robot’s collaboration performance over multiple days in different collaborative tasks and across diverse humans and scenes. Furthermore, we conduct extensive experiments to assess how well our simulated humans reflect real human behavior, particularly their ability to exhibit distinct, trait-driven patterns aligned with human profiles.

In summary, our main contributions are threefold:

- We present COOPERA, a novel HRC framework for continual, open-ended collaboration with humans who exhibit individual traits across long horizons.
- We develop a method to simulate humans with long-term behavior models driven by individual traits and habits.
- Within this framework, we introduce a benchmark and an approach that enables increasingly adaptive and personalized collaboration with humans over multiple days.

2 Related Work

Human-Robot Collaboration. Prior HRC work has largely focused on controlled lab settings [19, 10, 44, 59], where collaborative tasks are shared by both the human and the robot or narrowly defined. More recent research has expanded to complex household environments, requiring robots to infer human intentions from a single demonstration [50, 20, 64] or in an online fashion [51]. Subsequent works explore human intention inference using data from images [39] or simplified environments (e.g., 2D worlds) [5, 55, 70, 68], progressing to simulated real-world environments [20] and leveraging recent advances in VLMs. However, these approaches typically rely on predefined, closed-form representations of human intentions and tasks [14, 32, 2, 12, 71], and often ignore realistic human

behavior. Furthermore, collaboration is usually limited to fixed episodes with predefined task set. In contrast, our work considers open-ended and continual HRC, where humans spontaneously propose their actions based on environmental factors, and the collaboration persists across days.

Human Simulation. Most embodied AI works [6, 3, 16, 17, 60] assume that environmental changes are solely driven by a single robot [52]. Due to the challenges of real-human experiments (e.g., safety, scalability, cost), recent research has integrated deformable humans with plausible motion and appearance into robot simulation platforms [49, 52], enabling the study of safe and scalable HRC. However, these simulated humans focus only on motion feasibility, lacking the complexity and variability of real human behavior. Another research direction simulates humans with psychological traits and social interactions [76, 43, 46, 26, 64], but remains language-based and does not involve environmental interaction. In contrast, we simulate humans driven by psychological traits and habits, whose behavior is long-term and capable of interacting with their environment.

LLMs for Human Task Inference. One line of research treats human intentions as direct inputs and investigates how LLMs can interpret open-ended natural language instructions to generate structured robot plans [25, 24, 23, 73, 63, 38]. These works use techniques such as 3D scene graphs [56, 36, 8] to semantically ground high-level goals and decompose them into actionable subgoals, or incorporate human feedback [58, 33, 9] to quantify uncertainty and enable skill acquisition through interaction. In contrast, COOPERA takes a step further by aiming to let LLMs/VLMs infer personalized task plans, adapting to specific human traits and habits rather than general commonsense knowledge.

3 COOPERA: Continual, Open-Ended HRC Framework

Our goal is to enable the study of continual HRC in open-ended tasks. To that end, we investigate how a robotic agent can become more effective in assisting humans by learning from their behavior. Central to COOPERA are LLM-powered simulated humans driven by traits and long-term intentions that the robot can reason for effective collaboration, and a human feedback mechanism for improving collaboration over time. We first outline our framework and problem setup, detailing the collaboration settings we explore (Fig. 2). Then, we describe our approach of simulating humans driven by traits with long-term behaviors (Fig. 3). Finally, we propose a method to tackle our framework (Fig. 4).

3.1 Overview

In order to investigate HRC in a safe and reproducible manner, we consider a simulated human agent that interacts in a 3D household environment to achieve a set of high-level goals. These goals vary throughout the day and are driven by human traits and habits, as well as by the activities the human has done before. The robot’s goal is to assist the human in those tasks, without receiving explicit commands about the goal they should help with, or information about the human’s traits. Both human and robot have full knowledge of the environment. Each day is represented as 12 one-hour intervals covering the time from 9 am to 9 pm (the rest is treated as being asleep). At the beginning of each hour, the human proposes a high-level intention (e.g., leisure) and decomposes it into a sequence of tasks and executes them in the environment (e.g., watch TV on the sofa). As the human interacts in the environment, the robot has to infer the human’s goals and provide assistance. At the end of each day, the human provides feedback on the robot’s help, which is then used to improve the robot’s collaboration success in subsequent days.

Problem Setup. We define two types of collaboration with increasing difficulty and openness. **Collaboration type 1** is an open-ended variant of the Watch-and-Help challenge [50], where one intention (e.g., set up dinner table) is decomposed into 3 pick-and-place tasks (i.e., picking an object and placing it on a static object). For each intention, the robot is given a video of the human

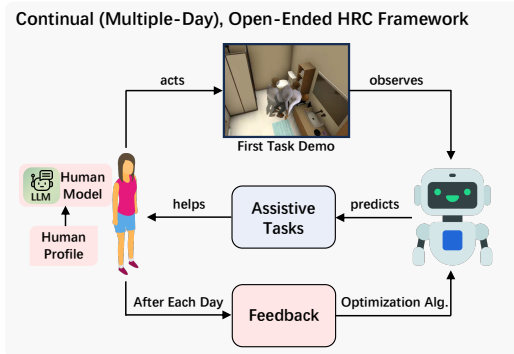


Figure 2: **COOPERA: Continual, open-ended human-robot collaboration framework.** The LLM-powered human proposes whole-day intentions and tasks, executed in the environment. As the robot observes the human actions, it predicts a set of tasks to assist them. After each day, the human provides feedback to the robot, enabling the robot to improve for subsequent days.

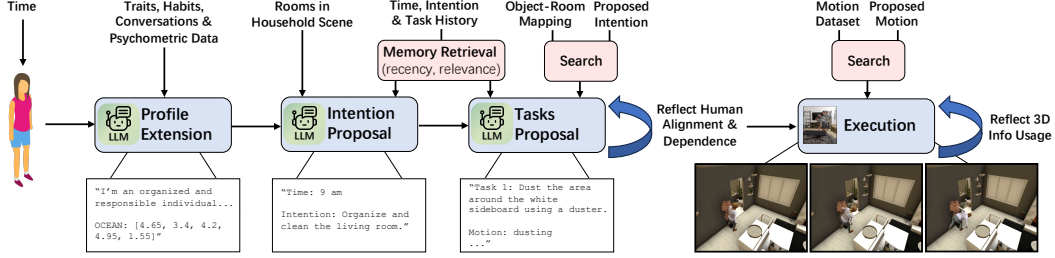


Figure 3: **Human Simulation Pipeline.** We seed the human-LLM with an extended profile. At each time of day, the human proposes an intention and decomposes it into tasks, aligning with profile traits and temporal dependence on intention/task history. LLM inputs are optimized with *Memory Retrieval* and *Search*, and robustness is enhanced via two rounds of *Reflexion*. This pipeline generates continuous, whole-day intentions and tasks executed in the environment with expressive whole-body motion. See Appendices C and F for details.

performing the first task and its textual description, and must infer and assist with the remaining tasks based on objects available in the scene. **Collaboration type 2** is more challenging and moves beyond pick-and-place: each intention (e.g., morning hygiene) is decomposed into 5 tasks involving free-form human motion while interacting with static objects (e.g., the human brushes teeth at the mirror). Unlike type 1, the robot is unconstrained by the scene and may propose any object it deems helpful (e.g., the robot offers toothbrush). It receives only the first task video with no textual guidance.

Evaluation Settings. We define four progressively challenging settings. **1) Same human, same scene:** The robot collaborates with the same human in the same scene over 5 consecutive days (5 days, 1 scene). **2) Same human, different scenes:** The robot collaborates with the same human across 5 different scenes, with a new scene each day (5 days, 5 scenes). **3) Different humans, same scene:** The robot collaborates with different humans in the same scene, rotating among Human 1, 2, and 3, each for one day, repeating this cycle three times in the same scene (9 days, 1 scene). **4) Different humans, different scenes:** The robot collaborates with different humans across multiple scenes, rotating through Human 1, 2, and 3 in the first scene, then repeating this sequence in the second and third scenes (9 days, 3 scenes). In **3)** & **4)**, we explore if knowledge gained from interacting with different humans improves future collaboration, despite fewer interaction days per human.

3.2 Simulating Humans

We aim to model humans who interact in the environment over long periods of time, act driven by their goals, preferences, and context, and who can react and provide feedback as a robot assists them. To achieve this, we propose a hierarchical model that combines LLMs and 3D human motion to simulate long-term, realistic human behaviors in indoor environments. First, our model generates a description of human traits describing their preferences and habits. Based on these traits, the environment, and the history of human actions, the model then generates a sequence of tasks for the human to perform. For every task, we use the environment information to generate human motions and interactions, providing a realistic demonstration of each task. Fig. 3 shows an overview of our design. Next, we describe in more detail each of the human simulation components.

Generating Human Traits. We use LLMs to generate personality traits that determine the human long-term behaviors. For this, we sample conversations from the Synthetic Human Dataset [26], containing dialogues between different humans, and prompt an LLM to generate a description of the human based on the conversation, inferring attributes such as their job, preferences or common activities. Inspired by [67, 77, 48], we also prompt the LLM to generate a vector measuring Big-5 human personality traits [18] (openness, conscientiousness, extroversion, agreeableness, and neuroticism), allowing us to measure the diversity across generated humans and how well the robot agents can infer the human’s personality from their interactions.

Whole-Day Intentions and Tasks. Given human traits, we generate long-term human behaviors, with tasks featuring temporal dependences within a day and diversity across days. *Temporal Dependence:* Given 3D environment information, we use an LLM to propose intentions for different times of day (e.g., 9 am: clean the living room). Next, we prompt the LLM to decompose the intentions into a sequence of inter-dependent tasks (e.g., dust the area around the white board, clean the counter). The LLM also receives the human’s intention and task history from previous hours, and is explicitly prompted to consider their inter-dependency. *Varying Distribution:* While humans with specific traits follow general routines, their daily behavior varies daily (i.e., Monday 9 am for cleaning, Tuesday 9

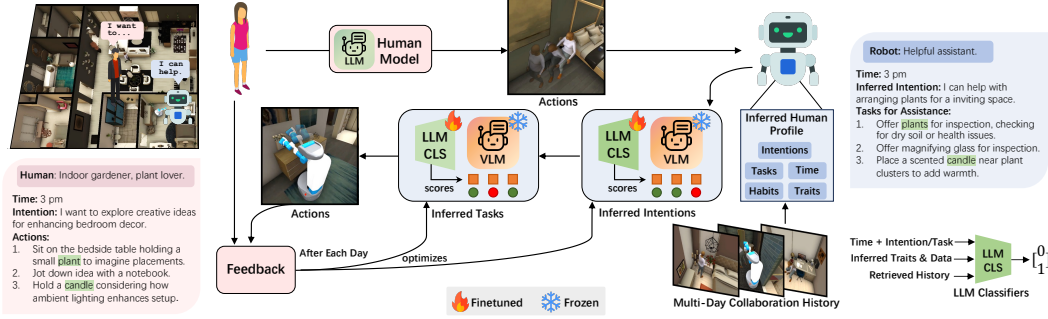


Figure 4: **Our approach for human assistance.** We decouple robot task inference into intention and task inference. By chaining VLM and classifier, the robot selects tasks aligned with the human’s traits and temporal context. It maintains a human profile inferred from collaboration history, which, combined with feedback, optimizes the robot-VLM via prompting and the classifiers via supervised learning. See Appendices D and G for details.

am for exercise). To model this, we reset the intention and task history at the start of each new day, setting a high temperature for the human-LLM to encourage diversity across days.

Expressive Whole-Body Motion. We simulate human agents physically using expressive 3D whole-body motions during task execution [31, 21, 42] by chaining motion sequences for each task. For pick-and-place tasks, the sequence includes walking, reaching and picking, walking again, then reaching and placing. For tasks involving free-form motion (e.g., sitting on a sofa), the human-LLM describes a free-form human motion that matches each task, using examples from our human motion dataset. The resulting sequence combines walking with the selected free-form motion.

Optimizing Long-Context Inputs. Our progressive prompting chain provides the human-LLM with substantial information at each stage, especially during the task proposal stage, where the 3D environment may contain hundreds of objects and the motion dataset includes thousands of data points. Additionally, as the day progresses, the intention and task history grows long (e.g., from 9 am to 9 pm, 13 intention sentences and dozens of task descriptions accumulate). Since LLMs struggle with long-context inputs [30, 35, 69], we introduce *Search* and *Memory Retrieval* mechanisms. *Search*: Given a query text and a list of texts, we return the top-K most relevant items based on semantic similarity. *Memory Retrieval*: We use recency and relevance scores to retrieve the top-K memories. Recency decays over time with a decay factor λ from the current time, and relevance is calculated by semantic similarity, similar to search. The final retrieval score is the product of both [46].

Self-Corrections. Given the complexity of our progressive prompting process and the LLM responses, even state-of-the-art models can make mistakes. Therefore, during the most complex task proposal stage, we perform two rounds of *Reflexion* [61, 74, 72] to identify and correct errors related to human traits, temporal dependencies, and object use within the 3D environment.

3.3 Instantiating COOPERA with an Assistive Agent

To study COOPERA, we propose an approach (Fig. 4) for continual HRC, enabling the robot to learn correlations between human intentions, tasks, traits, and temporal dependencies at each time of day.

At any given time, a human’s intentions/tasks can be viewed as meta-intentions/meta-tasks, encompassing a range of possible options due to the diversity of human behavior across days. Our solution decouples task inference into two stages: first inferring intentions, then identifying specific tasks. We capture the correct sets by chaining VLM to imagine multiple possible intentions/tasks and classifiers to score and filter them. Given observation (frames uniformly extracted from a video $V = [f_1, \dots, f_N]$) of the human’s first task, the robot-VLM generates an intention superset. For each positively classified intention by the intention classifier, the robot-VLM infers a set of possible tasks, forming a task superset. The task classifier then identifies the tasks most suitable for collaboration.

We optimize the robot-VLM through prompting and the binary classifiers via supervised learning. Using human feedback from the end-of-day discussion, the robot keeps tracks of a human profile by prompting robot-VLM to infer and summarize the human’s traits, habits, and psychometric data. This human profile, along with the retrieved history of intentions and tasks, is incorporated into the robot-VLM prompts and provided as input to the classifiers in the subsequent times and days. The robot-VLM and classifiers are optimized per day. Please see Fig. 4 for the input data format.

Table 1: **Evaluation of 1) human classification, 2) simulated human diversity, 3) human traits-psychometrics coherence, 4) temporal dependence, and 5) user studies.**

Classification (Acc $\uparrow$)		Diversity (SD $\uparrow$)	Coherence (R $\uparrow$)		Temporal Dependence		User Studies (Acc $\uparrow$)	
intention	task		aligned	mismatched	Acc $\uparrow$	F1 $\uparrow$	MCQ	Matching
0.995	0.830	0.939	0.342	-0.497	0.789	0.790	0.764	0.712

4 Experiments and Analysis

Within COOPERA, we first examine 1) if the central component, the simulated human model, reflects real human behavior and to what extent. 2) We then introduce the benchmark setup (baselines, evaluation metrics) and explore if our proposed approach leads to more personalized robot assistance over multiple days compared to baselines. 3) Subsequently, we analyze the real-world applicability of our framework. 4) Finally, we evaluate the effectiveness of each module through ablation studies.

4.1 Framework Implementation

Environment and Scene. We use *Habitat 3.0* [52] as the robot simulation platform and *HSSD* [28] as the 3D environment, which includes 18,656 static objects across diverse scenes in style and size. Since the original HSSD includes only static objects, we develop a systematic approach to create dynamic scenes by making small objects from specific categories (e.g., decor, kitchenware) movable. We also sample 20 dynamic objects from the *YCB Dataset* [7] and place them in contextually appropriate locations (e.g., a mug on a bedside table) using Habitat’s built-in tools. Dynamic scenes are initialized at the start of each episode. Across days, Habitat tracks object locations as the human and robot interact with the environment, allowing them to maintain updated environment knowledge. We select 5 scenes with varying of rooms (4–11), static objects (51–140), and dynamic objects (33–94). All scenes provide enough space for the human and Fetch robot [15] to navigate. Please see Appendix B for more details.

Human Dataset. For modeling unique humans, we use the *SPC: Synthetic-Persona-Chat Dataset* [26], a fully synthetic dataset that includes hundreds of short user profiles along with their conversations and compute psychometric data (details in Appendix C). We use *Motion-X* [31] and *AMASS* [41] as the human motion dataset. We generate 10 human profiles.

Training and Inference. We use open-source models for interpretability and benchmarking value. For simulating humans, we use Llama-3.1-8B [13] with temperature 0.7. For search and memory retrieval, we use MiniLM-L6-v2 [66] with a decay factor $\lambda = 0.95$, retrieving the top 3 intentions and top 5 tasks. For the assistive agent, we use Llama-3.2-11B [13] as the robot-VLM. Classifiers are finetuned on Mistral-7B-Instruct-v0.2 [27] using LoRA [22] in instructional format to output binary yes/no. We train on 3 NVIDIA A10 GPUs (24GB RAM). Please see Appendix D for more details.

4.2 Analysis of Human Simulation

Distinct Simulated Humans. We examine if simulated humans with distinct Big-5 traits exhibit machine-identifiable features. Each of 10 humans is placed in 5 scenes, living 20 days per scene. We aggregate daily intentions and tasks into one data point per human. Two 10-way BERT-large-uncased classifiers [11] are finetuned—one for intentions (10 epochs), one for tasks (20 epochs) with train-test split 0.8:0.2, learning rate $5e-6$, and tested on an unseen scene. As shown in Table. 1, task classification is harder than intention classification, as intentions align more with human traits, but tasks (e.g., drinking water) may correspond to multiple intentions (e.g., leisure, exercise).

Diverse Simulated Humans. We assess diversity by standard deviation (SD) of Big-5 traits (1–5 scale) across 10 simulated humans. We compute per-trait SD and take average. From Table. 1, the high SD exceeds the typical 0.7–0.9 range in real-world distributions [62], validating diversity.

Human Traits and Psychometrics Coherence. In our main approach, the robot-VLM infers Big-5 traits throughout the day based on the human’s intention and task history. Using the final scores at the end of collaboration, we assess coherence with ground truth via Pearson correlation [48, 4], and introduce a one-step mismatch for comparison. The significant drop in correlation for mismatched pairs (Table. 1) confirms alignment between inferred traits and psychometric data, demonstrating the LLM’s ability to interpret human psychology from behavior.

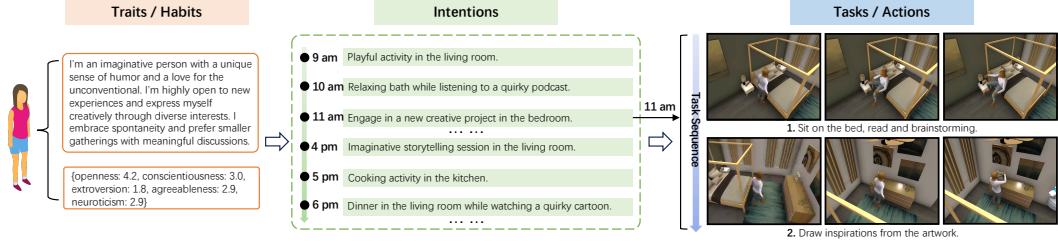


Figure 5: **Qualitative examples of full-day intentions and tasks** proposed by a human with specific human traits and psychometric data.

Temporal Dependence in Human Behavior. We study how current-hour intentions depend on prior hours via a next-sentence prediction task: given three earlier intentions (e.g., 9–11 am), the 12 pm intention is used as the positive example, with other-time intentions as negatives. We evaluate on 10 simulated humans, each living for 20 days in 5 scenes. BERT-large-uncased [11] is trained for 20 epochs (learning rate 5e-6). Results in Table. 1 confirm temporal dependence.

User Studies. We conduct two user studies (25 participants each) to assess: 1) Whether real humans can identify the same simulated human across days and scenes, and 2) Whether real humans can distinguish simulated humans with varying traits and Big-5 scores. For 1), we sample full-day intentions and tasks of 10 simulated humans across 2 days and 2 scenes (4 samples/human), then construct 10 multiple-choice questions showing a human profile and three behavior options (1 correct, 2 distractors). For 2), we sample full-day behaviors from 10 simulated humans with distinct traits, and ask participants to match trait descriptions to the corresponding full-day intentions and actions. Results in Table. 1 show higher accuracy in identifying the same simulated human than in distinguishing between different simulated humans. This discrepancy likely arises because multiple-choice tasks provide explicit answer options, reducing ambiguity, whereas trait-based matching requires deeper reasoning about personality-behavior relationships, making it more challenging.

Alignment with Real-Human Behavior. We study how simulated human intentions align with real-human intentions. We recruit six participants who provide personality traits and psychometric data, and record their daily intentions over five days. Using these traits, we prompt the LLM to generate simulated intentions for the same time span. To assess alignment, we aggregate both sets into single paragraphs (removing time formatting like “9am: ...” to avoid inflated structural similarity) and compute semantic similarity using SBERT (all-mpnet-base-v2) [57] and OpenAI embeddings (text-embedding-3-small) [1]. We compare against: 1) prompting without human profile (generic) and 2) mismatched LLM-human intention pairs (mismatched). From Table 2, both baselines yield moderate similarity (~ 0.5), as sentence encoders assign partial similarity to structurally similar content. In contrast, aligned pairs achieve much higher scores, indicating strong alignment between simulated and real human intentions. SBERT slightly outperforms OpenAI embeddings, likely due to its sentence-level training objective.

Table 2: **Semantic alignment** between simulated and real-human intentions.

	Generic	Mismatched	Main
SBERT $\uparrow$	0.554	0.523	0.810
OpenAI Emb. $\uparrow$	0.537	0.543	0.772

Qualitative Results. We present examples of full-day intentions and tasks proposed by a human with specific human traits and psychometric data in Fig. 5.

4.3 Analysis of Continual, Open-Ended HRC

Since COOPERA involves long-term, open-ended tasks that requires the robot reasoning over human traits and temporal context, we construct baselines using standard LLM/VLM-based approaches adapted for task inference. These baselines reflect commonly used paradigms in open-world robot planning [8, 29].

Baselines. 1) **Direct Prompting:** The robot proposes a single intention from visual input and decomposes it into tasks. The robot-VLM is optimized solely via prompting with retrieved intention/task history. 2) **Direct Finetuning:** The robot brain is finetuned to directly output a single intention and decompose it into tasks. 3) **Oracle:** The robot is given the ground-truth human intention and decomposes it into tasks. 4) **Random:** Intention and task classifiers are removed; all proposed intentions and tasks are accepted without validation. 5) **Intention Agnostic:** The robot directly

predicts and filters tasks without first inferring intentions. **6) Human & Context Agnostic:** The classifiers do not learn the correlation between human traits and intentions/tasks or the temporal dependence between previous and current intentions/tasks. They only learn the relationship between the current time and the intentions/tasks.

Evaluation Metrics. We assess the assistive agent’s performance using F1-based success rate across three methods, ensuring a comprehensive evaluation from simulation to real-world perspectives, incorporating prior HRC approaches [51, 8]. **1) Predicate-based:** Tasks are executed and evaluated by predicate functions with success based on object class matches rather than instance matches, following Watch-and-Help [50]. **2) LLM-based:** Given the ground truth human intention and predicted tasks, the human-LLM judges whether each task fulfills the intention (binary yes/no), and F1 is computed over these labels. **3) Human verification:** Same as 2) but evaluated by real human users. In our main method, the robot-VLM generates a task superset, which is filtered by a task classifier assigning yes/no labels used for F1 computation. For baselines without a task classifier (e.g., Direct Prompting), all predicted tasks are treated as positive. Please see Appendices C and D for details on predicate construction and LLM evaluation.

Setup. We follow the evaluation settings in Section 3. Setting 1 evaluates 5 humans across 2 scenes; Setting 2 evaluates 10 humans; Setting 3 spans 3 distinct scenes; and Setting 4 spans 9 humans.

Analysis of Assistive Performance.

We analyze two aspects: 1) Within-day improvement—does the robot’s collaboration success increase throughout the day by learning temporal dependencies between human intentions and actions? 2) Across-day improvement—does collaboration become more successful and personalized over multiple days, using end-of-day feedback? From Fig. 6 (a), our method achieves the highest within-day improvement. In contrast, prompting, random, and oracle exhibit little to no improvement, or even decline. We hypothesize that these methods do not

benefit from learning human intentions, which are highly correlated with human traits and temporal context. This finding aligns with our human classification experiment in Section 4.2. From Fig. 6 (b), our method shows the strongest improvement across days, second to oracle. The minimal gain in prompting and finetuning highlights the challenge of varying human behavior across days, as these methods tend to establish a 1-to-1 mapping between time and human intentions/tasks. Prompting relies heavily on collaboration history, while finetuning prioritizes the highest probability training data, limiting adaptability to varying human behaviors.

Out-of-Domain Generalization. We study **1) Scene generalization:** can the robot personalize collaboration with a human in an unseen scene after interacting in other scenes? **2) Human generalization:** can the robot collaborate effectively with a new human after training with others? For 1), we use models finetuned from setting 2 on four scenes and evaluate on a fifth, unseen scene with the same human. The baseline is the robot’s average performance during its initial

interaction in the unseen scene, without finetuning on the previous scenes, averaged over 10 humans. For 2), we use models finetuned from setting 3 on three humans and evaluate on collaboration with a fourth unseen human. The baseline is the robot’s unadapted performance when first interacting with the new human, without finetuning on the previous three, averaged over 5 new humans. Result in Table. 3 show that generalizing to a new human is much harder than to a new scene. This is likely due to greater variability in human behaviors. While the robot can learn shared patterns across humans, effective collaboration with a new human requires adaptation to fine-grained, person-specific traits.

Qualitative Results. We show how the robot improves collaboration within a day by inferring more correct tasks for assistance, along with a visualization of HRC at a specific time in Fig. 7.

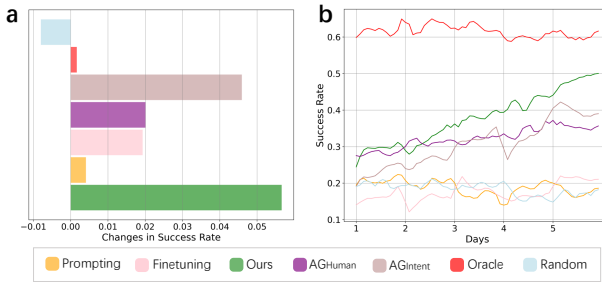


Figure 6: **Evaluation of changes in robot success rate (predicate-based):** (a) within a single day and (b) across multiple days. See Appendix E for a detailed breakdown of the results and additional analysis.

Table 3: **Generalization performance.** We report the average success rate (predicate-based).

	Baseline	Finetuned
Scene	0.269	0.465
Human	0.258	0.343

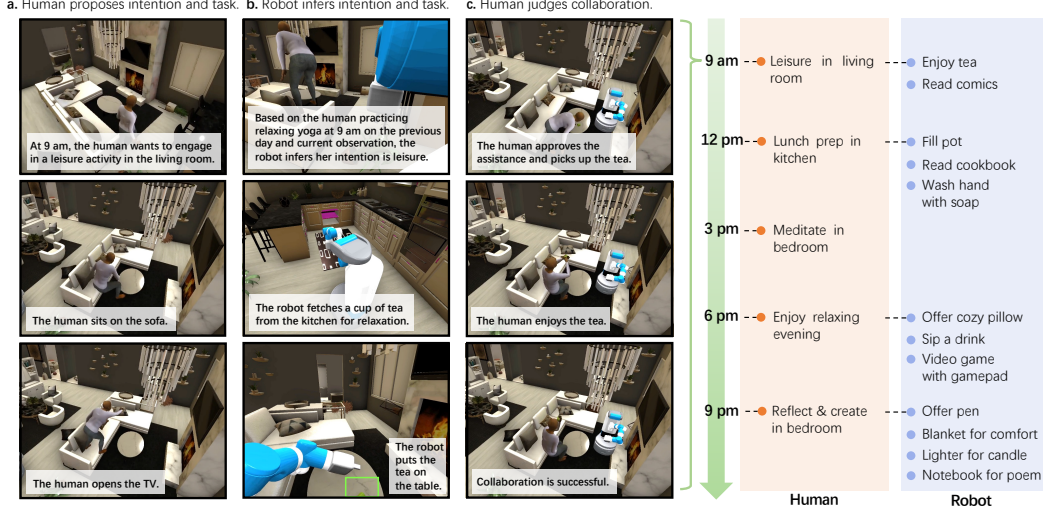


Figure 7: **Qualitative examples of successful human-robot collaboration within one day.** The red column displays human intentions, while the blue column shows the robot’s correctly inferred tasks for assistance.

4.4 Analysis of Real-World Applicability

The ultimate goal of COOPERA is to develop robot agents that assist real humans by adapting to their preferences over long-term. Yet, fully real-world experiments pose ethical, safety, and cost issues (requiring a real human and physical robot to interact in a household over multiple days). To validate applicability of COOPERA under real-world conditions, we take three complementary approaches. Please see Appendix E for detailed results, qualitative examples, and additional analysis.

Human Verification. We validate if predicate-based and LLM-based evaluations align with human verifications by sampling one episode per setting from our main method. From Table. 4, the low L1 indicates strong correlation among all three metrics, supporting our framework’s real-world applicability. This justifies our use of VLMs/LLMs for predicting human actions in both baselines and main method, as their reasoning closely resembles human decision-making. Also, in open-ended settings with large state spaces, LLMs/VLMs serve as effective reasoning modules due to their generalization abilities [8, 29].

Table 4: **Correlation between predicates, LLM, and human evaluations (rows 1–2):** L1 ↓, averaged over collaboration types. **Offline real-human and human-in-the-loop collaboration (row 3–4):** predicate-based success rate (SR ↑), averaged over the final day.

	Setting 1	Setting 2	Setting 3	Setting 4
Predicate vs. Real-Human (L1 ↓)	0.091	0.091	0.085	0.120
LLM vs. Real-Human (L1 ↓)	0.077	0.080	0.077	0.075
Offline Real Human (SR ↑)	0.498	0.471	0.426	0.322
Human-in-the-Loop (SR ↑)	0.488	0.467	0.431	0.349

Collaborating with Offline Real Humans. Real humans exhibit greater behavioral dynamics and emergent decisions due to temporary factors (e.g., plans, mood, weather). We recruit six participants who provide personality traits and psychometric data, and record their daily intentions over five days. An LLM decomposes these into tasks in HSSD scenes, where the robot collaborates across all settings. Results in Table 4 show performance comparable to simulated humans (Table. 6 row 4), despite increased dynamics.

Human-in-the-Loop. Six real humans replace the LLM and collaborate with the assistive agent. They are shown retrieved object and motion sets and select which to interact with based on their intention. For ease of implementation and usability, participants input responses as text rather than using a keyboard to control the simulated agent. Results in Table 4 indicate that real human collaborators do not make the task more difficult and can, in some cases, lead to higher success rates compared to offline or simulated humans.

4.5 Ablation Studies

Human Simulation. **1) Removing human profile extension:** To explore whether our method yields the most distinct simulated humans, we remove simulated conversations and profile extension, prompting LLM only with the original short trait paragraph. We evaluate by finetuning and measuring

Table 5: **Ablation study on human simulation.** The effects of removing profile extension on human classification and using single-shot intention proposal on temporal dependence.

	Removing Profile Extension (Acc $\uparrow$)		All-Day Intention Proposal	
	intention cls.	task cls.	Acc $\uparrow$	F1 $\uparrow$
Removed	0.950	0.800	0.751	0.740
Ours	0.995	0.830	0.789	0.790

Table 6: **Ablation study on assistive agent:** success rate (predicate-based) $\uparrow$ averaged over the last day.

	Setting 1	Setting 2	Setting 3	Setting 4
No Traits	0.481	0.443	0.239	0.206
No Context	0.452	0.414	0.408	0.299
Changing Backbone	0.487	0.424	0.362	0.310
Ours (main)	0.505	0.465	0.439	0.344

classifiers accuracy (Section 4.2). **2) Single-shot human intention proposal:** We test if our pipeline design maximizes temporal dependence in human behavior. Instead of proposing intentions hour by hour with access to intention/task history and using Reflexion, we remove these and generate all intentions at once. We evaluate via next-intention prediction (Section 4.2). Results in Table. 5 confirm the effectiveness of the profile extension module and the overall pipeline design.

Assistive Agent. **1) Removing human traits inference.** We examine the importance of learning human traits by removing robot’s inference of human traits or Big-5 scores from past intentions and tasks, preventing classifiers from learning their correlation. **2) Removing temporal context learning:** We assess the impact of learning temporal dependence between human intentions and tasks by preventing classifiers from using past intentions/tasks when predicting the current one. **3) Changing the robot brain backbone:** We replace the robot-VLM from Llama-3.2-11B [13] to LLaVA-1.6-Mistral-7B [34]. Using different VLMs for human and robot reduces alignment and tests robustness. From Table 6, removing trait inference significantly reduces success rate in settings 3 and 4 involving multiple humans, as the robot struggles to distinguish them. The learning of time-based context benefits all settings. Despite smaller model size, the robot still achieves reasonable success.

5 Conclusion

We introduce **COOPERA**, a framework for continual, open-ended HRC. We propose a human model to generate long-term human behaviors driven by personality traits, a benchmark, and a method to assist humans under by predicting their long-term intentions. Our framework opens up exciting directions for future work, such as using communication to better infer human traits or build agents that can perform proactive assistance (e.g. arranging a house before the start of the day based on preferences). We hope that this work can promote future research on building agents that can work over long time horizons and adapt to human preferences.

Limitations. Despite compelling HRC performance, COOPERA currently focuses on single-human settings, leaving multi-human collaboration for future work. While evaluations are primarily conducted in simulation, we validate sim-to-real transfer through real-human routines and interactions. Our method uses skill primitives compatible with standard robot platforms (e.g., Fetch), making it readily transferable to real hardware.

References

- [1] Open AI. New embedding models and api updates. <https://openai.com/index/new-embedding-models-and-api-updates/>, 2024.
- [2] Stefano V. Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, pages 66–95, 2018.
- [3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018.
- [4] Danny Azucar, Davide Marengo, and Michele Settanni. Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, pages 150–159, 2018.
- [5] Chris L. Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B. Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 2017.
- [6] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*, 2020.
- [7] Berk Çalli, Arjun Singh, Aaron Walsman, Siddhartha S. Srinivasa, Pieter Abbeel, and Aaron M. Dollar. The YCB object and model set: Towards common benchmarks for manipulation research. In *International Conference on Advanced Robotics*, pages 510–517. IEEE, 2015.
- [8] Matthew Chang, Gunjan Chhablani, Alexander Clegg, Mikael Dallaire Cote, Ruta Desai, Michal Hlavac, Vladimir Karashchuk, Jacob Krantz, Roozbeh Mottaghi, Priyam Parashar, Siddharth Patki, Ishita Prasad, Xavier Puig, Akshara Rai, Ram Ramrakhya, Daniel Tran, Joanne Truong, John M. Turner, Eric Undersander, and Tsung-Yen Yang. Partnr: A benchmark for planning and reasoning in embodied multi-agent tasks. In *International Conference on Learning Representations*, 2025.
- [9] Jinghong Chen, Guangyu Yang, Weizhe Lin, Jingbiao Mei, and Bill Byrne. On extending direct preference optimization to accommodate ties. *arXiv preprint arXiv:2409.17431*, 2024.
- [10] Kerstin Dautenhahn. Socially intelligent robots: dimensions of human-robot interaction. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, pages 679–704, 2007.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019.
- [12] Yuqing Du, Stas Tiomkin, Emre Kiciman, Daniel Polani, Pieter Abbeel, and Anca D. Dragan. Ave: Assistance via empowerment. In *Advances in Neural Information Processing Systems*, 2020.
- [13] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra,

- Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [14] Alan Fern, Sriraam Natarajan, Kshitij Judah, and Prasad Tadepalli. A decision-theoretic model of assistance. *Journal of Artificial Intelligence Research*, 50:71–104, 2014.
- [15] Fetch Robotics. Fetch, 2020.
- [16] Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, Kuno Kim, Elias Wang, Michael Lingelbach, Aidan Curtis, Kevin T. Feiglis, Daniel Bear, Dan Gutfreund, David D. Cox, Antonio Torralba, James J. DiCarlo, Josh Tenenbaum, Josh H. McDermott, and Dan Yamins. Threedworld: A platform for interactive multi-modal physical simulation. In *Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.
- [17] Théophile Gervet, Soumith Chintala, Dhruv Batra, Jitendra Malik, and Devendra Singh Chaplot. Navigating to objects in the real world. *Science Robotics*, 2023.
- [18] L R Goldberg. An alternative “description of personality”: the big-five factor structure. *Journal of Personality and Social Psychology*, pages 1216–1229, December 1990.
- [19] Michael A. Goodrich and Alan C. Schultz. Human-robot interaction: A survey. *Foundations and Trends in Human-Computer Interaction*, pages 203–275, 2007.
- [20] Moritz A. Graule and Volkan Isler. GG-LLM: geometrically grounding large language models for zero-shot human activity forecasting in human-aware task planning. In *International Conference on Robotics and Automation*, pages 568–574. IEEE, 2024.
- [21] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Conference on Computer Vision and Pattern Recognition*, pages 5142–5151. IEEE, 2022.
- [22] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.
- [23] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147, 2022.
- [24] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Tomas Jackson, Noah Brown, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning*, pages 1769–1782, 2022.
- [25] Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. Do as I can, not as I say: Grounding language in robotic affordances. In *Conference on Robot Learning*, pages 287–318, 2022.
- [26] Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. Faithful persona-based conversational dataset generation with large language models. In *Findings of the Association for Computational Linguistics*, pages 15245–15270, 2024.

- [27] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
- [28] Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, and Manolis Savva. Habitat synthetic scenes dataset (HSSD-200): an analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In *Conference on Computer Vision and Pattern Recognition*, pages 16384–16393. IEEE, 2024.
- [29] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li Erran Li, Ruohan Zhang, Weiyu Liu, Percy Liang, Li Fei-Fei, Jiayuan Mao, and Jiajun Wu. Embodied agent interface: Benchmarking llms for embodied decision making. In *Advances in Neural Information Processing Systems*, 2024.
- [30] Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhui Chen. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*, 2024.
- [31] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. In *Advances in Neural Information Processing Systems*, 2023.
- [32] Chang Liu, Jessica B. Hamrick, Jaime F. Fisac, Anca D. Dragan, J. Karl Hedrick, S. Shankar Sastry, and Thomas L. Griffiths. Goal inference improves objective and perceived performance in human-robot collaboration. In *International Conference on Autonomous Agents & Multiagent Systems*, pages 940–948. ACM, 2016.
- [33] Haokun Liu, Yaonan Zhu, Kenji Kato, Atsushi Tsukahara, Izumi Kondo, Tadayoshi Aoyama, and Yasuhisa Hasegawa. Enhancing the llm-based robot manipulation through human-robot collaboration. *IEEE Robotics and Automation Letters*, pages 6904–6911, 2024.
- [34] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Conference on Computer Vision and Pattern Recognition*, pages 26286–26296. IEEE, 2024.
- [35] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, pages 157–173, 2024.
- [36] Yuchen Liu, Luigi Palmieri, Sebastian Koch, Ilche Georgievski, and Marco Aiello. DELTA: decomposed efficient long-term robot task planning using large language models. *arXiv preprint arXiv:2404.03275*, 2024.
- [37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.
- [38] Kai Lu, Chenyang Ma, Chiori Hori, and Diego Romeres. Kitchenvla: Iterative vision-language corrections for robotic execution of human tasks. In *International Conference on Robotics and Automation, Workshop on Safe Vision-Language Models for Robotics*, 2025.
- [39] Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and Andrew Markham. Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors. In *Proceedings of the Conference on Neural Information Processing Systems*, 2024.
- [40] Chenyang Ma, Xinchu Qiu, Daniel Beutel, and Nicholas Lane. Gradient-less federated gradient boosting tree with learnable learning rates. In *Proceedings of the 3rd Workshop on Machine Learning and Systems*, pages 56–63, 2023.

- [41] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: archive of motion capture as surface shapes. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5441–5450. IEEE, 2019.
- [42] Aymen Mir, Xavier Puig, Angjoo Kanazawa, and Gerard Pons-Moll. Generating continual human motion in diverse 3d scenes. In *International Conference on 3D Vision*, pages 903–913. IEEE, 2024.
- [43] Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, and Zhongyu Wei. Agentsense: Benchmarking social intelligence of language agents through interactive scenarios, 2024.
- [44] Stefanos Nikolaidis, Ramya Ramakrishnan, Keren Gu, and Julie A. Shah. Efficient model learning from joint-action demonstrations for human-robot collaborative tasks. In *International Conference on Human-Robot Interaction*, pages 189–196. ACM, 2015.
- [45] OpenPsychometrics. The big five personality test, n.d.
- [46] Joon Sung Park, Joseph C. O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Symposium on User Interface Software and Technology*, pages 2:1–2:22. ACM, 2023.
- [47] Claudia Pérez-D’Arpino, Can Liu, Patrick Goebel, Roberto Martín-Martín, and Silvio Savarese. Robot navigation in constrained pedestrian environments using reinforcement learning. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1140–1146. IEEE, 2021.
- [48] Heinrich Peters and Sandra C. Matz. Large language models can infer psychological dispositions of social media users. *arXiv preprint arXiv:2309.08631*, 2023.
- [49] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Conference on Computer Vision and Pattern Recognition*, pages 8494–8502, 2018.
- [50] Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B. Tenenbaum, Sanja Fidler, and Antonio Torralba. Watch-and-help: A challenge for social perception and human-ai collaboration. In *International Conference on Learning Representations*, 2021.
- [51] Xavier Puig, Tianmin Shu, Joshua B. Tenenbaum, and Antonio Torralba. NOPA: neurally-guided online probabilistic assistance for building socially intelligent home assistants. In *International Conference on Robotics and Automation*, pages 7628–7634. IEEE, 2023.
- [52] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander Clegg, Michal Hlavac, So Yeon Min, Vladimir Vondrus, Théophile Gervet, Vincent-Pierre Berges, John M. Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A co-habitat for humans, avatars, and robots. In *International Conference on Learning Representation*, 2024.
- [53] Xinchu Qiu, Heng Pan, Wanru Zhao, Chenyang Ma, Pedro Porto Buarque de Gusmao, and Nicholas Donald Lane. Efficient vertical federated learning with secure aggregation. In *Federated Learning Systems (FLSys) Workshop@ MLSys 2023*, 2023.
- [54] Xinchu Qiu, Heng Pan, Wanru Zhao, Chenyang Ma, Pedro PB Gusmao, and Nicholas D Lane. vfedsec: Efficient secure aggregation for vertical federated learning via secure layer. *arXiv preprint arXiv:2305.16794*, 2023.
- [55] Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew M. Botvinick. Machine theory of mind. In *International Conference on Machine Learning*, pages 4215–4224, 2018.

- [56] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian D. Reid, and Niko Sünderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. In *Conference on Robot Learning*, pages 23–72, 2023.
- [57] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *International Joint Conference on Natural Language Processing*, pages 3980–3990, 2019.
- [58] Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In *Conference on Robot Learning*, pages 661–682, 2023.
- [59] Leonel Dario Rozo, Sylvain Calinon, Darwin G. Caldwell, Pablo Jiménez, and Carme Torras. Learning physical collaborative robot behaviors from human demonstrations. *IEEE Transactions on Robotics*, pages 513–527, 2016.
- [60] Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. Vint: A foundation model for visual navigation. In *Conference on Robot Learning*, pages 711–733, 2023.
- [61] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.
- [62] Christopher J Soto and Oliver P John. The next big five inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, pages 117–143, 2017.
- [63] Andrew Szot, Max Schwarzer, Harsh Agrawal, Bogdan Mazouze, Rin Metcalf, Walter Talbott, Natalie Mackraz, R. Devon Hjelm, and Alexander T. Toshev. Large language models as generalizable policies for embodied tasks. In *International Conference on Learning Representations*, 2024.
- [64] Yanming Wan, Yue Wu, Yiping Wang, Jiayuan Mao, and Natasha Jaques. Infer human’s intentions before following natural language instructions. *arXiv preprint arXiv:2409.18073*, 2024.
- [65] Chen Wang, Claudia Pérez-D’Arpino, Danfei Xu, Li Fei-Fei, Karen Liu, and Silvio Savarese. Co-gail: Learning diverse strategies for human-robot collaboration. In *Conference on Robot Learning*, pages 1279–1290. PMLR, 2022.
- [66] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, 2020.
- [67] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Conference of the Association for Computational Linguistics*, pages 5635–5649, 2019.
- [68] Sarah A. Wu, Rose E. Wang, James A. Evans, Joshua B. Tenenbaum, David C. Parkes, and Max Kleiman-Weiner. Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, pages 414–432, 2021.
- [69] Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. Effective long-context scaling of foundation models. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4643–4663, 2024.
- [70] Fengyu Yang and Chenyang Ma. Sparse and complete latent organization for geospatial semantic segmentation. In *Conference on Computer Vision and Pattern Recognition*, pages 1809–1818, 2022.

- [71] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. *Advances in Neural Information Processing Systems*, pages 8081–8103, 2022.
- [72] Guangyu Yang, Jinghong Chen, Weizhe Lin, and Bill Byrne. Direct preference optimization for neural machine translation with minimum Bayes risk decoding. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 391–398, 2024.
- [73] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.
- [74] Mert Yükeşgönül, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. Textgrad: Automatic "differentiation" via text. *arXiv preprint arXiv:2406.07496*, 2024.
- [75] Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485*, 2023.
- [76] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Annual Meeting of the Association for Computational Linguistics*, pages 270–278, 2020.
- [77] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, 2024.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: We conducted extensive empirical experiments to demonstrate it.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We discussed limitations in the Conclusion.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[NA\]](#)

Justification: We did not have theoretical assumptions.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We have thoroughly discussed implementations details in the main paper and Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We did not provide code with the submission because of internal regulations within the authors' organizations but will release it after acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We discussed these in the Experiments and Appendix with a lot of details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: In our experiments, large time and computation cost is needed to do this.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We discussed it in the Experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We followed the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Our work does not have broader social impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No such concern.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All properly cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: All well documented.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: We do not have large-scale crowdsourcing experiments.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[NA\]](#)

Justification: This work does not have such.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We explained in great details in framework and experiment sections.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

Appendix for *COOPERA*

A Overview

This Appendix includes: 1) more details about how we construct dynamic HSSD scenes and statistics, 2) additional details of how we build, train, and evaluate the simulated human and the assistive agent, 3) additional HRC results and analysis, including both quantitative metrics and qualitative examples, along with comparisons to human verification of our main method and baselines across collaboration types and settings, and 4) prompt details.

B Additional Details of Scenes

B.1 Dynamic Habitat Synthetic Scenes Dataset Construction

We make certain objects in the HSSD [28, 40, 53, 54] scenes dynamic by checking if they are supported by any structure and their object super categories. For objects that do not have support, we classify them as follows:

dynamic_categories = [trashcan, decor, dining ware, plant, electronics, animate object, apparel, liquid container, kitchen ware, tray, bathroom accessory, gym equipment, toy, wearable]

static_categories = [storage furniture, support furniture, seating furniture, floor covering, lighting, sleeping furniture, bathroom fixtures, mirror, large kitchen appliance, large appliance, kitchen bathroom fixture, vehicle, heating cooling, medium kitchen appliance, display, arch, curtain, small kitchen appliance]

B.2 Scene Statistics

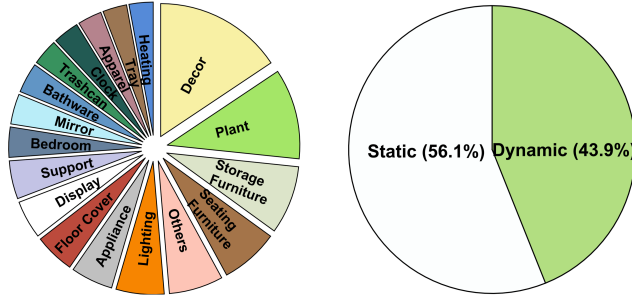


Figure 8: **Distribution of object categories** within the constructed dynamic HSSD scenes.

Our selected 5 scenes feature varying number of rooms (4–11), static objects (51–140), and dynamic objects (33–94). We present the distribution of 18 representative object categories (out of 32) in Fig. 8. The large number and diversity of objects and categories enable humans to propose a wide range of open-ended tasks.

B.3 Scene Summarization and Visualizations

We summarize 3D environment information in each scene as a text-based dictionary, which is used as input to the simulated humans. Specifically, we extract the bounding boxes of rooms and map each object to its corresponding room, forming an object-room mapping in the format of object_ID: [object_name, room]. For example:

mapping = {'Nemo Kepler Pendant, Black': [125, 'corridor'], 'AquaVive stordoucheset Kila met kraan': [123, 'main bathroom'], 'Uttermost Marlow Chandelier': [118, 'bathroom of bedroom 1'], 'Eisa Pendant': [114, 'main bedroom'], 'CAR - SUV': [12, 'garage'], 'Nolan Upholstered King Bed': [109, 'main bedroom'], ... }

We show visualizations of the five scenes used in COOPERA in Fig. 9.

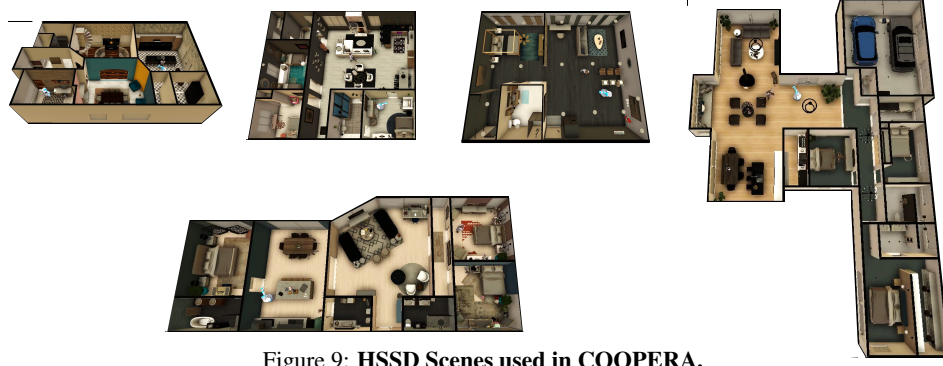


Figure 9: HSSD Scenes used in COOPERA.

C Additional Details of Human Simulation

C.1 Psychometric Data Computation

Since SPC dataset lacks psychometric data, we derive Big-5 OCEAN scores by prompting the LLM to 1) directly infer the scores [48] and 2) complete the Big-5 personality test [18, 45] and compute scores based on the formula. We then take a majority vote across five inference trials, using bins of 0.5 on a scale of 1-5.

C.2 3D Motion

For human simulation, free-form motion data is formatted in SMPL-X, enabling detailed control over whole-body motions, including facial expressions and finger articulation. To integrate this data into the human simulation pipeline of Habitat 3.0 [52], we remap the format of Motion-X [31] data, as illustrated below.

global root orientation: SMPL-X[:, :3]
body: SMPL-X[:, 3:3+63]
finger articulation: SMPL-X[:, 66:66+90]
yaw pose: SMPL-X[:, 66+90:66+93]
face expression: SMPL-X[:, 159:159+50]
global body position: SMPL-X[:, 209:209+100]
global body position: SMPL-X[:, 309:309+3]
body shape: SMPL-X[:, 312:]

C.3 Feedback

Please refer to Section F for details on our designed feedback mechanism for the robot’s predicted assistive tasks. We emphasize that while we structure the feedback as binary answers paired with reasoning, it can be easily adapted for other learning algorithm by modifying the prompt.

C.4 Additional Qualitative Examples

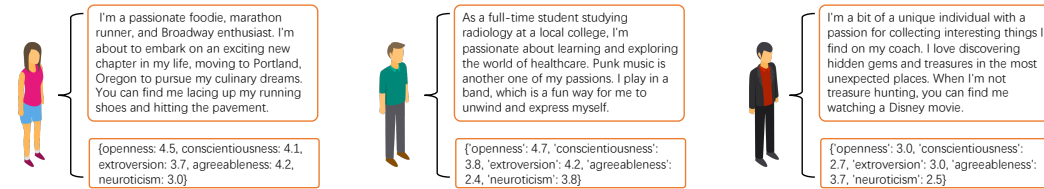


Figure 10: Additional examples of simulated human profiles psychometric data.

We present additional examples of simulated human profiles along with their corresponding psychometric data in Fig. 10.

D Additional Details of Building and Evaluating an Assistive Agent

D.1 Navigation and Movement

To navigate the scene, the agent uses classical motion planning for path calculation, with a stopping threshold determined by the object’s Axis-Aligned Bounding Box (AABB). For the Fetch robot’s arm movements, inverse kinematics (IK) is applied to calculate the end effector (EE) position based on the 3D location of the target object. Robot actions are implemented using primitives from Habitat’s task_action registry, combining navigation and EE control to form a full pick-and-place pipeline: 1) MoveEEAction moves the EE via Cartesian steps with IK; 2) PickObjIdAction grasps objects using a snap mechanism within a distance threshold; 3) PlaceObjIdAction releases objects via desnap; and 4) ResetEEAction resets the EE to a default pose.

D.2 Training

Task videos from Habitat 3.0 [52] are resized to 1024×768 and input to Llama-3.2-11B [13] (our robot-VLM) using Llama’s default settings. Classifiers are finetuned on Mistral-7B-Instruct-v0.2 [27] using LoRA [22] (rank 8, dropout 0.2, alpha 16; targets: q, k, v, o) in an instructional format to output binary yes/no. We train for 5 epochs using AdamW [37] (lr $1e-5$, weight decay 0.01), with batch size 1 and gradient accumulation of 4 steps, across 3 NVIDIA A10 GPUs (24GB RAM).

For training the robot’s intention and task binary classifiers via instructional finetuning of an LLM, we structure the input data in the following format:

```

### Instruction:
Considering the human’s profile, traits, temporal dependence on past behaviors, and the current time,
determine if it is likely or unlikely that this human will: ... Respond with 'Yes' or 'No'.

### Input:
Human Profile.
Big Five Traits.
Previous Relevant Intentions.
Previous Relevant Tasks.
Current Time.

### Response:

```

D.3 Predicate Construction

Collaboration Type 1	Time: 6 pm Intention: Enjoy a playful and imaginative dinner preparation in the kitchen, experimenting with new recipes and flavors.	Task 1: Place the "tuna_fish_can" on the "KITCHEN AID ARTISAN MIXER RED" for setup. Task 2: Place the "mustard_bottle" on the "Edelweiss desk, ash/white" in the living room. Task 3: Place the "fork" on the "Edelweiss desk, ash/white" in the living room.	ON(food, kitchen electricals) ON(food, table) ON(kitchen items, table)
Collaboration Type 2	Time: 9 am Intention: Engage in a morning workout routine in the living room.	Task 1: Start with a warm-up by performing yoga stretches near the "Marina Slipcover Sofa" to prepare the body for more intense exercises. Task 2: Perform squats with dumbbells near the "Wendover Art # Morning Lake View I" for strength training. Task 3: Transition to cardio by doing jumping jacks near the "Wendover Art # Off the Path" to elevate heart rate. Task 4: Engage in core exercises by doing planks near the "APPLE iMac 5K 27" to strengthen abdominal muscles. Task 5: Cool down with a stretching routine near the "Marina Slipcover Sofa" to relax muscles and prevent injury.	NEED(yoga mat) NEED(dumbbells) NEED(jump rope) NEED(exercise mat) NEED(towel)

Figure 11: Examples of predicate construction.

For collaboration type 1, we map objects to Habitat 3.0 and the YCB dataset’s predefined semantic categories, allowing evaluation via exact match. For collaboration type 2, where the robot offers objects from a magic box, we compute semantic similarity between the robot’s predicted tasks/objects for assistance and the human’s desired objects, using a threshold of 0.6. Fig. 11 illustrates examples of predicate construction.

E Additional Experiments and Details

E.1 Results Breakdown and Analysis

In Fig. 12, we show detailed predicate-based and LLM-based evaluations in both collaboration types across all collaboration settings. Overall, LLM-based evaluations yield higher scores than predicate-based ones, particularly in settings 3 & 4, which are more challenging. This suggests that predicate-based evaluations are more rigorous, requiring exact matches for collaboration type 1 and high semantic similarity for collaboration type 2. We highlight several key findings and analyze the performance of each baselines in detail below.

Temporal Fluctuation Patterns Across Days. In Fig. 12, we observe a consistent temporal pattern across days: performance tends to dip around midday and recover toward the end of the day, and improve further on the following day. We attribute this to two primary factors. First, it reflects how human routines are structured, both in reality and in our simulation. Humans generally engage in more predictable activities in the morning and evening (e.g., hygiene, eating, relaxing), which are easier for the robot to infer and assist with. In contrast, midday behavior tends to be more diverse and more strongly influenced by individual traits. Since our human simulation pipeline samples intentions and tasks based on both traits and time, this increases the diversity of midday actions, making inference and alignment more difficult for the robot and resulting in a temporary performance drop. Performance recovers as routine behaviors re-emerge in the evening or the following morning. Importantly, despite these fluctuations, overall performance improves across days. Second, the extent of this fluctuation increases with task difficulty, as defined in our framework (Section 3.1). In Setting 1 (same human, same scene), the robot benefits from consistent exposure to both the user and environment, resulting in relatively small midday dips. In Setting 2 (same human, different scenes), unfamiliar object layouts introduce grounding challenges that increase midday variability. In Setting 3 (different humans, same scene), rotating between users interrupts personalization, making intention interpretation more difficult—particularly around midday when behavior is less routine. Finally, Setting 4 (different humans, different scenes) combines both challenges, resulting in the largest fluctuations as the robot must generalize across both human and spatial contexts.

Performance Analysis: Main Method vs. Baselines. We evaluate our method against six baselines: 1) Direct Prompting, 2) Direct Fine-tuning, 3) Oracle, 4) Random, 5) Intention Agnostic, and 6) Human & Context Agnostic. Fig. 12 shows that our method achieves both the highest adaptation trend over time and the highest final success rate (second only to Oracle) by explicitly modeling the correlation between human intentions/tasks, traits, and temporal dependencies. As the day progresses, the robot accumulates interaction history and adapts its behavior to the human preferences. 1) Direct Prompting shows minimal improvement as it depends on retrieved interaction history without learning the human preferences. 2) Direct Finetuning performs slightly better but still struggles due to its rigid mapping from inputs to intentions/tasks, making it biased toward frequent training examples and overlooking the varying distribution of human behavior. 3) Oracle serves as an upper-bound performance reference by receiving ground-truth human intentions. While this provides a significant advantage, its performance is static by design. This is because Oracle bypasses the core challenge of accurately mapping visual observations of humans performing detailed tasks to high-level intentions, a fundamental problem in human-robot collaboration, even in closed-set scenarios. Moreover, since the correct intention is provided upfront, Oracle has no need to learn temporal dependencies between human intentions (e.g., an intense workout at 11 am typically leads to lunch preparation at 12 pm). 4) Random degrades over time due to the absence of learning or validation. 5) Intention Agnostic shows moderate improvement but lags behind our method, as it skips intention inference and thus loses contextual information, which is problematic since temporal dependencies between intentions are typically stronger than between tasks (validated in Section 4.2). 6) Human & Context Agnostic captures only superficial time-task correlations and ignores human traits or temporal context, leading to weak within-day gains.

E.2 Human Verification Breakdown and Analysis

For human verification, we present additional results across three baselines and our main method, analyzing how human evaluations align with predicate-based (Table. 7) and LLM-based (Table. 8) assessments in a detailed breakdown. Due to cost constraints, we use a subset of data (one episode per setting), resulting in 32 episodes across the four methods. We recruit eight human evaluators, each

Table 7: **Breakdown of correlation between predicates and human evaluations.** We report the L1 ↓.

	Collaboration 1				Collaboration 2			
	Setting 1	Setting 2	Setting 3	Setting 4	Setting 1	Setting 2	Setting 3	Setting 4
Random	0.070	0.134	0.067	0.065	0.073	0.079	0.088	0.112
Oracle	0.072	0.098	0.084	0.065	0.086	0.102	0.104	0.108
Human & Context Agnostic	0.062	0.121	0.066	0.071	0.073	0.109	0.080	0.100
Main	0.092	0.096	0.094	0.087	0.090	0.085	0.076	0.153
Average	0.074	0.112	0.078	0.072	0.080	0.094	0.087	0.118

Table 8: **Breakdown of correlation between LLM and human evaluations.** We report the L1 ↓.

	Collaboration 1				Collaboration 2			
	Setting 1	Setting 2	Setting 3	Setting 4	Setting 1	Setting 2	Setting 3	Setting 4
Random	0.031	0.054	0.066	0.044	0.068	0.091	0.084	0.110
Oracle	0.103	0.096	0.087	0.043	0.071	0.069	0.091	0.094
Human & Context Agnostic	0.045	0.058	0.076	0.071	0.084	0.089	0.079	0.108
Main	0.063	0.075	0.081	0.073	0.090	0.085	0.072	0.076
Average	0.061	0.071	0.078	0.058	0.078	0.084	0.082	0.097

assessing four randomly assigned episodes. The results show that the gap between predicate-based evaluation and human verification is larger than that between LLM-based evaluation and human verification. This is because LLMs/VLMs reason and make judgments for assistance in a way that aligns more closely with human preferences and decision-making, whereas predicate-based evaluation relies on strict matching criteria that may not fully capture nuanced human reasoning.

E.3 Qualitative Examples of Offline Real Human

We present qualitative examples of offline real humans in Fig. 13. Compared to simulated humans (Fig. 5), real humans exhibit greater behavioral dynamics and emergent decisions due to temporary factors (e.g., plans, mood, weather).

E.4 Analysis of Human-in-the-Loop Collaboration

Intuitively, one might expect lower performance when collaborating with real humans due to their inherent variability and spontaneity. However, results in Table 4 indicate that real human collaborators do not make the task more difficult and can, in some cases, lead to higher success rates compared to offline or simulated humans. Through discussions with participants, we identify a likely explanation. Unlike the offline setting where participants record their routines throughout an entire day (12 one-hour intervals covering the time from 9 am to 9 pm), the HITL experiments take place in a short, controlled session. As a result, participants do not act spontaneously hour-by-hour but instead recall and follow their typical weekly routines from memory, which tend to be more consistent. This reduced behavioral variability allows the robot to personalize more quickly, and further validates the behavioral diversity modeled in our simulated humans.

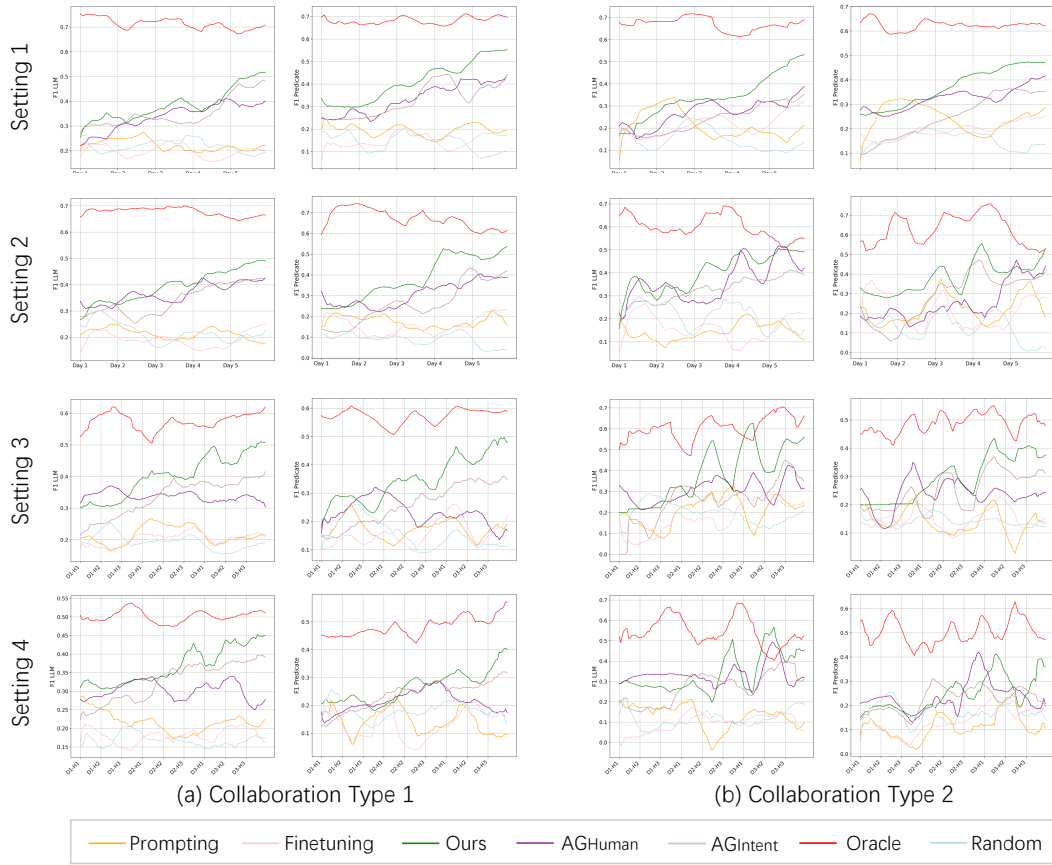


Figure 12: **Breakdown results of HRC.** We compare our main method with the baselines across days.

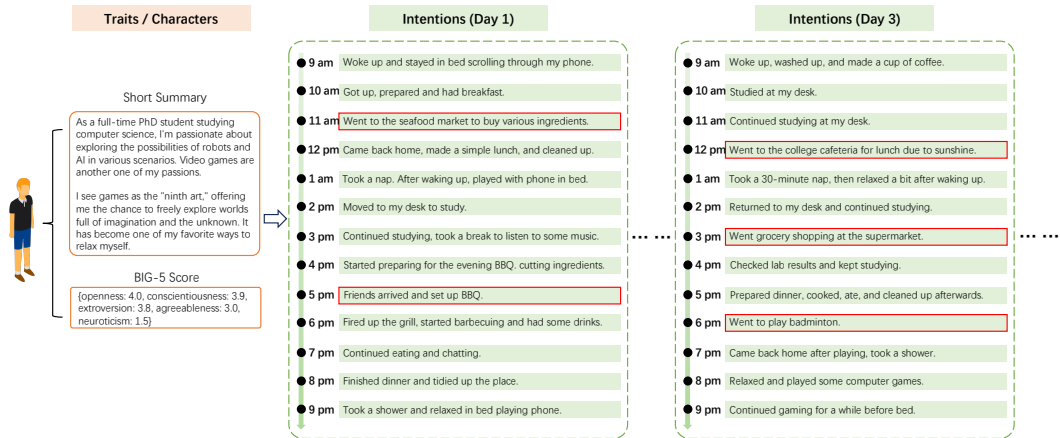


Figure 13: **Qualitative examples of full-day intentions** recorded by a real human with specific human traits and psychometric data. Red boxes highlight emergent behaviors due to temporary factors (e.g., plans, mood, weather).

F Prompt Details of the Simulated Human

We show exact prompts for LLMs in simulated human behavior.

Human Profile Summary and Extension.

Input:

1. Human 1 profile.
2. Human 2 profile.
3. Conversation between Human 1 and Human 2.

Instruction:

Summarize and reasonably expand Human 1's profile into a first-person self-introduction based on their initial profile and past conversations with other humans. Provide a detailed description covering, if presented, the person's job, hobbies, daily activities, food preferences, social life, physical activity, entertainment preferences, travel habits, personal values, goals and aspirations, stress and coping mechanisms, technological use, cultural interests, health and wellness, community involvement, education, financial habits, and personal style.

Human Psychometric Data Computation (LLM Inference).

Input:

1. Human profile.

Instruction:

Infer Big Five personality traits (scale 1-5, float) based on the provided human profile. Write in the following format: {'openness': a, 'conscientiousness': b, 'extroversion': c, 'agreeableness': d, 'neuroticism': e}

Human Psychometric Data Computation (Big-5 Personality Test).

Input:

1. Human profile.

Instruction:

Based on the provided human profile, complete the Big-5 personality test. In the questions below, for each statement 1-50 mark how much you agree with on the scale 1-5, where 1=disagree, 2=slightly disagree, 3=neutral, 4=slightly agree and 5=agree.

Questions:

1. Am the life of the party.
2. Feel little concern for others.
-
50. Am full of ideas.

Human Intention Proposal.

Input:

1. Current time.
2. A list of rooms in the house (ignore small spaces like closets).
3. Your Big Five scores (scale 1-5) and human profile.
4. Most relevant human intentions proposed at previous times (ignore if empty—this means it's the first intention of the day).
5. Most relevant human tasks proposed at previous times (ignore if empty—this means it's the first task of the day).

You are a human living in the house. Propose your intention at the current time.

Instructions:

1. Intention must align with your Big 5 scores, reflect all aspects of the profile, and be diverse yet reasonable based on the house layout and available objects.

2. Intention must be high-level and either human-centric (e.g., hygiene, sport, leisure) or room-centric (e.g., clean, organize, set-up). Do not mention specific objects.
3. Intention must have temporal dependence but be non-repetitive with the previous intentions and tasks.
4. Intention should be within the house.
5. All objects are rigid and cannot deform, disassemble, or transform.

Write in the following format. Do not output anything else:

Time: xxx am/pm (e.g., 9 am)

Intention: basic descriptions.

Reason_human: detailed descriptions of why it follows your Big 5 scores and profile.

Reason_intentions: detailed descriptions of why it has temporal dependence with the previous, relevant intentions at [list of time].

Reason_tasks: detailed descriptions of why it has temporal dependence with the previous, relevant tasks at [list of time.id].

Human Task Proposal.

Input:

1. The proposed intention at current time.
2. A dict mapping rigid, static objects to their IDs and rooms.
3. Your Big Five scores (scale 1-5).
4. Most relevant human intentions proposed at previous times (ignore if empty—this means it's the first intention of the day).
5. Most relevant human tasks proposed at previous times.ids (ignore if empty—this means it's the first task of the day).

You are a human living in the house.

Instructions:

1. Break down the intention into 5 tasks for collaboration with a robot.
2. Task types:
 - Type 1: Creative, reasonable free-form human motion interacting or approaching a fixed, static object (static objects cannot be moved) with an object in hand provided by the robot (e.g., sit on sofa with TV remote control in hand, wipe table with tissue in hand, squat with dumbbell in hand near rug).
3. For interacting with fixed, static objects, use only objects from the given static object dict. For objects in hand, a robot will provide them.
4. Both interacting and inhand objects must be specified (cannot be none).
5. Tasks should be continuous and logical, and align with your Big 5 scores and profile.
6. Tasks must have temporal dependence with the intentions and tasks at previous times.
7. Free-form motion should be diverse. Examples: sampled_motion_list. Feel free to propose others.
8. All objects are rigid and cannot deform, disassemble, or transform.

Write in the following format. Do not output anything else:

Time: xxx am/pm

Intention: basic descriptions.

Tasks:

1. Thought: detailed descriptions of the task. Reason_human: why it aligns with your Big 5 scores and profile. Reason_intentions: how it depends on previous, relevant intentions at [list of time]. Reason_tasks: how it depends on previous, relevant tasks at [list of time.id]. Act: [type: 1, inter_obj_id: real int, inter_obj_name: xxx, inhand_obj_name: yyy, motion: free-form motion]
2. ...

Human Reflection (Profile).

Input:

1. The proposed intention at current time.
2. A dict mapping rigid, static objects to their IDs and rooms.

3. Your Big Five scores and human profile.
4. Most relevant human intentions proposed at previous times (if empty, ignore it—this means it's the first intention of the day).
5. Most relevant human tasks proposed at previous times.ids (if empty, ignore it—this means it's the first intention of the day).

Your task is to check if the temporal dependence and human profile are strictly followed in each task, and revise to make better if necessary.

Instructions:

1. Tasks should be continuous and logical, and align with your Big 5 scores and profile.
2. Tasks must have temporal dependence with the previous intentions and tasks, with detailed explanation mentioning previous intentions and tasks explicitly.
3. For interacting with fixed, static objects, use only objects from the given static object dict. For objects in hand, a robot will provide them.

Write in the following format. Do not output anything else:

Time: xxx am/pm

Intention: basic descriptions.

Reflect Each Task:

1. no mistake or change made.
2. ...

Revised Tasks:

1. Thought: detailed descriptions of the task. Reason_human: why it aligns with your Big 5 scores and profile. Reason_intentions: how it depends on previous, relevant intentions at [list of time]. Reason_tasks: how it depends on previous, relevant tasks at [list of time.id]. Act: [type: 1, inter_obj_id: real int, inter_obj_name: xxx, inhand_obj_name: yyy, motion: free-form motion]
2. ...

Human Reflection (3D Info).

Input:

1. The proposed intention at current time.
2. A dict mapping rigid, static objects to their IDs and rooms.

Your task is to check if the instructions are strictly followed in each task, and revise to make better if necessary.

Instructions:

1. Break down the intention into 5 tasks for collaboration with a robot.
2. Task types:
 - Type 1: Creative, reasonable free-form human motion interacting or approaching a fixed, static object (static objects cannot be moved) with an object in hand provided by the robot (e.g., sit on sofa with TV remote control in hand, wipe table with tissue in hand, squat with dumbbell in hand near rug).
3. For interacting with fixed, static objects, use only objects from the given static object dict (exact name). For objects in hand, a robot will provide them.
- 4 Both interacting and inhand objects must be specified. Importantly, they cannot be none.
5. Free-form motion should be diverse. Examples: sampled_motion_list. Feel free to propose others.
6. All objects are rigid and cannot deform, disassemble, or transform.

Write in the following format. Do not output anything else:

Time: xxx am/pm

Intention: basic descriptions.

Reflect Each Task:

1. no mistake or change made.
2. ...

Revised Tasks:

1. Thought: detailed descriptions of the task. Reason_human: why it aligns with your Big 5 scores and profile. Reason_intentions: how it depends on previous, relevant intentions at [list of time]. Reason_tasks: how it depends on previous, relevant tasks at [list of time.id]. Act: [type: 1, inter_obj_id: real int, inter_obj_name: xxx, inhand_obj_name: yyy, motion: free-form motion]
2. ...

Feedback.

Input:

1. Human intention at current time.
2. Current human tasks.
3. Robot-inferred tasks to enhance comfort with offered objects (if any).

You are the human. Decide if the robot's assistance align with your needs.

Instructions:

1. Assess if each robot task supports the human tasks and intention. The robot's task doesn't need to be an exact match but should be relevant in purpose, context, or object categories. Use common reasoning to decide if it helps meet your needs.
2. Consider each robot thought and object individually against the human tasks. Approve it if it meets any one of the human tasks; sequence does not matter.
3. Be fair in your judgment—avoid being too generous or too harsh.
4. Respond with yes/no for each, followed by an explanation. Ensure items are in a list.

Write in the following format. Do not output anything else:

Tasks: [yes, no, ...]

Reasons_tasks:

1. ...

G Prompt Details of the Assistive Agent

We show exact prompts for VLMs in building the assistive agent.

Intention Discovery.

Input:

1. Sequence of images showing human motion from your and human's perspectives.
2. Current time.
3. Inferred Big Five personality scores (ignore if empty—this means it's your first collaboration with this human).
4. Inferred human profile (ignore if empty—this means it's your first collaboration with this human).
5. Most relevant human intentions discovered at previous times (ignore if empty—this means it's the first intention of the day).
6. Most relevant human tasks discovered at previous times.ids (ignore if empty—this means it's the first task of the day).

You are a robot assisting a human. Identify 5 possible human intentions based on the current time and visual observations.

Instructions:

1. Map the observed human motion to 5 possible high-level intentions at the current time (without mentioning the specific motion).
2. Intention must align with human Big 5 scores and reflect all aspects of the profile, and be diverse yet reasonable based on the house layout and available objects.
3. Intention must be high-level and either human-centric (e.g., hygiene, sport, leisure) or room-centric (e.g., clean, organize, set-up). Do not mention specific objects.

4. Intention must have temporal dependence but be non-repetitive with the intentions and tasks at previous times in the input.

Write in the following format. Do not output anything else: Time: xxx am/pm

Intention 1: basic descriptions.

Reason_human: detailed descriptions of why it follows the Big 5 scores and profile.

Reason_intentions: detailed descriptions of why it has temporal dependence with the previous, relevant intentions at [list of time].

Reason_tasks: detailed descriptions of why it has temporal dependence with the previous, relevant tasks at [list of time.id].

Reason_vis: detailed descriptions with respect to the visual cues.

...

Task Discovery.

Input:

1. Human intention at current time.
2. A dict mapping rigid, static furnitures to their IDs and rooms.
3. Inferred Big Five personality scores (ignore if empty—this means it's your first collaboration with this human).
4. Most relevant human intentions discovered at previous times (ignore if empty—this means it's the first intention of the day).
5. Most relevant human tasks discovered at previous times.ids (ignore if empty—this means it's the first task of the day).

You are a robot assisting a human.

Instructions:

1. Break down the intention into 5 tasks.
2. Task type: For each human task, provide one small, handable object from a magical box. Furnitures in the dict are for room understanding and cannot be used.
3. Tasks should be continuous and logical, and align with your Big 5 scores and profile.
4. Tasks must have temporal dependence with the intentions and tasks at previous times.
5. All objects are rigid and cannot deform, disassemble, or transform.

Write in the following format. Do not output anything else:

Time: xxx am/pm

Intention: basic descriptions.

Tasks:

1. Thought: detailed descriptions of the task. Reason_human: why it alignes with your Big 5 scores and profile. Reason_intentions: how it depends on previous, relevant intentions at [list of time]. Reason_tasks: how it depends on previous, relevant tasks at [list of time.id]. Act: [obj_name: xxx]
2. ...

Traits Inference.

Input:

1. Human intentions at previous times (ignore if empty—this means it's your first inference).
2. Human tasks at previous times.ids.
3. Human profile (ignore if empty—this means it's your first inference).

Task: Mimic this human by:

1. Inferring Big Five personality traits (scale 1-5, float) based on the provided intentions and task.
2. Summarizing the human profile (i.e., preferences/habits) based on the intentions and tasks within three sentences. Revise the existing human profile if necessary.

Write in the following format. Do not output anything else:

Scores: {'openness': a, 'conscientiousness': b, 'extroversion': c, 'agreeableness': d, 'neuroticism': e}

Profile: ...
Reasons_ocean: explain each ocean.
Reasons_profile: explain the profile.