A Framework for Continual and Open-Ended Human-Robot Collaboration

Chenyang Ma[1] Kai Lu[1] Ruta Desai[2†]
Xavier Puig[2†] Andrew Markham[1†] Niki Trigoni[1†]
[1]University of Oxford [2]FAIR, Meta

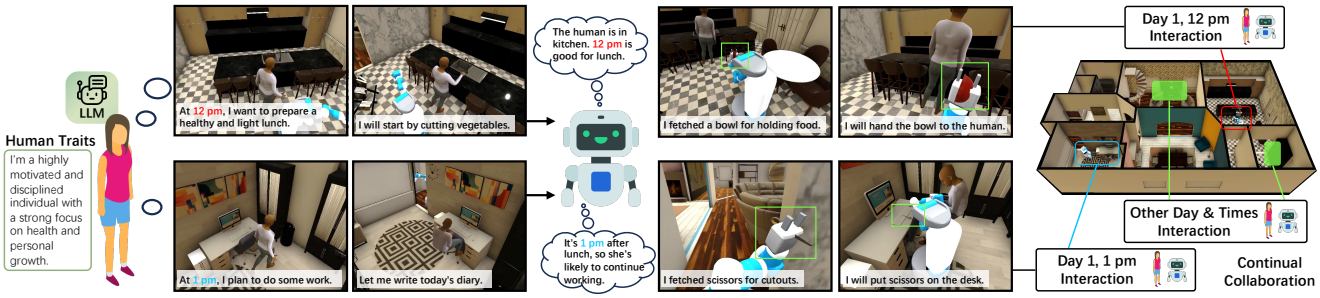

Figure 1. **Continual human-robot collaboration for open-ended tasks over multiple days.** Our framework entails an approach to simulate traits-driven humans with long-term, whole-day behaviors within robot simulation platform. We also propose a method for the robot to personalize collaboration in such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

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 a novel framework for continual, open-ended human-robot collaboration (HRC), 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 HRC in different types of collaborative tasks across various time-scales. Using this framework, we propose an approach to personalize the robot's collaborative actions by learning human traits and context-dependent intents. Experiments validate the realism of our simulated humans 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. Such agent should not only be able to fetch a cup to drink coffee, but also learn that someone may prefer the coffee cooler in the morning, but stronger in the afternoon, heating up water accordingly. Or it should also anticipate when a human will want to sit and relax, cleaning the couch before it is asked to. 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 reason about the human across long periods of time, adapting their behavior to provide better assistance.

Over recent years, several works have made significant advances in developing agents that can assist humans in household tasks [36, 39, 41, 47, 53], leveraging simulation environments to study human-robot collaboration 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 address these challenges, we present a novel human-robot collaboration (HRC) framework designed to facilitate continual, open-ended collaboration and adaptation with humans in complex household environments (Fig. 1). Our framework is designed around two core aspects that are necessary to study long-term HRC:

Traits-Driven Human Modeling. To represent realistic HRC scenarios in simulation, we need to model humans that can interact in the environment over long periods of time. We model such humans following 3 principles: **1)** Humans interact in the environment driven by goals or intentions, that

unfold over one day, reflecting long-term, evolving patterns. For instance, after setting the dinner table at 6 pm, a person might shift to watching TV at 7 pm. **2**) These goals and intentions should be open-ended and contextualized. Rather than following rigid, predefined tasks, humans spontaneously propose their intentions and tasks based on the environment such as the objects available or time of the day. **3**) These goals and intentions are in turn driven by psychological traits and habits, which inform the tasks they may do throughout the day. These traits also may be different for different humans and thus result in unique behaviors for different humans even within similar environments (e.g., a person may prefer to start the day reading a book whereas another human will prefer to clean the house).

Continual Collaboration. As the human interacts in the environment, we need a way to measure whether the robot is providing effective assistance over time. We also need a way to provide feedback to the robot, so that 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 day's end, the human communicates with the robot and provides feedback to help improve the robot's collaboration success rate for subsequent days.

Our framework presents unique challenges which are often overlooked in existing HRC benchmarks. First, robot agents need to reason not only about the state of the environment and the current human behavior, but also about why the human is behaving in such a way, so that they can assist more effectively in the future. 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 will perform different tasks depending on the time in 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 study this challenging framework, we propose a method, called *human imitation* which leverages a Vision-Language Model (VLM) to predict, given observations of the human behavior, the goals their are trying to achieve, and proposes actions to assist them in such goals. This approach 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 within our HRC framework over multiple days in different collaboration types and across diverse settings. Furthermore, we conduct extensive experiments to validate the realism of our simulated human and their ability to exhibit distinct, traits-driven behaviors that align with human profiles.

In summary, our main contributions are threefold:

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

2. Related Work

Human-Robot Collaboration. Prior work in HRC community has primarily been restricted to controlled lab environments [8, 17, 33, 43], where the goals of collaborative tasks are shared by both the human and the robot, or are defined within a narrow scope. More recent research has extended this setting to larger, more complex household environments, where human intentions must be inferred by observing a single demonstration [18, 39, 46] or in an online manner [40]. Subsequent works explore human intention inference using data from images [29] and toy or simplified environments (e.g., 2D worlds) [5, 42, 50], progressing to simulated real-world environments [18] and leveraging recent advances in VLMs. However, these approaches typically rely on predefined, closed-form representations of human intentions and tasks [2, 10, 12, 26], without accounting for realistic human behaviors. Furthermore, collaboration is usually limited to a fixed episode defined by a predetermined task set. In contrast, our work considers open-ended and continual human-robot collaboration, where humans spontaneously propose their actions based on environmental factors, and the collaboration persists across days.

Human Simulation. Most works in the embodied AI literature [3, 6, 14, 15, 44] operate under the assumption that changes to the environment are solely driven by the actions of a single robotic agent [41]. As early efforts stemmed from the challenges of real-human experiments (e.g., safety, scalability, cost) in developing agents capable of collaborating with humans, recent research has explored how to integrate deformable humans with plausible motion and appearance within robot simulation platforms [38, 41], allowing both robots and humans to interact with the environment. However, these simulated humans only consider motion feasibility, lacking the complexity and variability of real human behavior. Another line of research explores how to simulate human agents with psychological traits and their interactions in social contexts [21, 32, 35, 46, 55]. Yet, these works are primarily language-based and do not involve direct environmental interaction. In contrast, we investigate how to simulate humans driven by psychological traits and habits, whose behavior is long-term and capable of interacting with their environment.

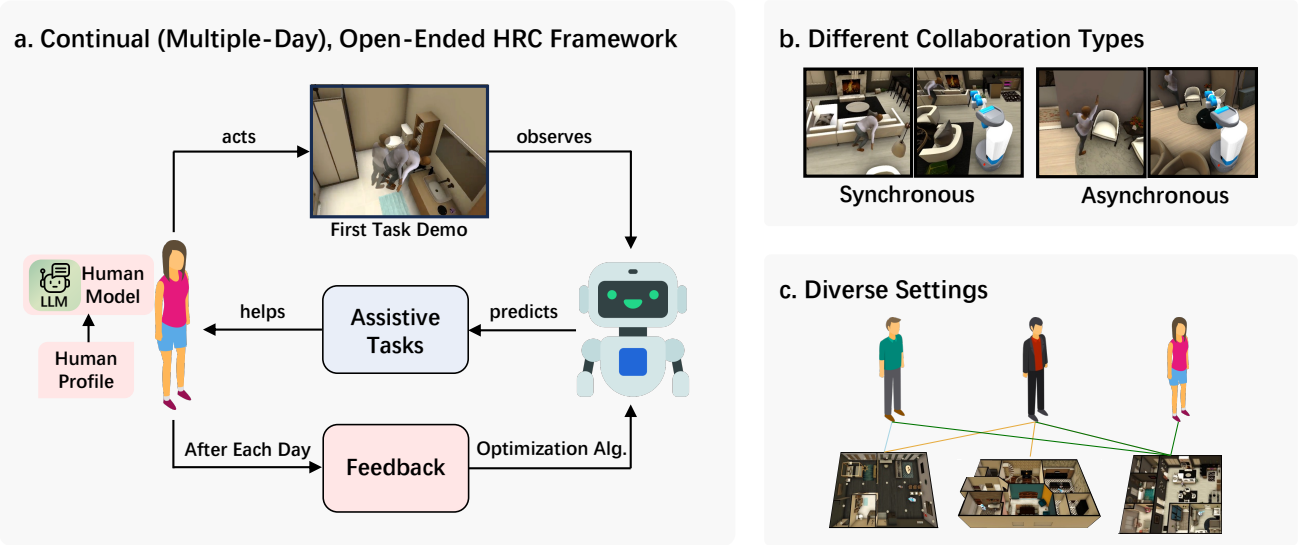

Figure 2. **(a) Continual, open-ended human-robot collaboration framework.** The LLM-powered human proposes whole-day intentions and tasks, which are executed in the environment. As the robot observes the human actions, it predicts a set of tasks to assist them. After each day, the human provides feedback to the robot, enabling the robot to improve for subsequent days. We study collaboration under (b) synchronous and asynchronous settings, and (c) across different humans and environments.

3. Continual, Open-Ended HRC Framework

Our goal is to enable continual HRC in open-ended tasks. To that end, we aim to investigate how a robotic agent can become more effective in assisting humans by learning from their behavior. Central to our framework are LLM-powered simulated humans driven by traits and long-term intentions that the robot can reason about for effective collaboration and a human feedback mechanism enabling improvement in collaboration over time. We first detail our framework and the problem setup of various collaboration settings that we study using the framework (Fig. 2). Then, we describe our approach of simulating humans driven by traits with long-term behaviors (Fig. 3). Finally, we introduce evaluation metrics and compare our proposed method with representative baselines on HRC over multiple days (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 goals or intentions. These intentions vary throughout the day and are driven by psychological traits and habits. The robot's goal is to assist the human in those tasks, without receiving explicit commands about the goal they should help with. Both the human and robot have holistic knowledge of the environment, but the robot lacks specific knowledge about the human. At each time point within the chosen time intervals in a day (e.g., from 9 am to 9 pm, in one-hour intervals), the human proposes an intention and decomposes it into Q tasks, which are executable in the environment. An intention represents a high-level, possibly vague plan (e.g., leisure), while a task is a specific, actionable item within that intention (e.g., watch TV on the sofa). Akin to Watch-and-Help challenge [39], the robot predicts the possible tasks to provide assistance by observing the human's first task execution. At the end of each day, a discussion period takes place where the human communicates with the robot, providing feedback on the human's true intentions and tasks and if the collaborations were successful. This feedback could be leveraged to improve the robot's collaboration success rate for subsequent days.

Problem Setup. We define two collaboration types with increasing levels of difficulty and openness. **Collaboration type 1** is an open-ended variant of the Watch-and-Help challenge [39]. In this type, one intention (e.g., set up dinner table) 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 provided with a video showing the human successfully performing the first task, along with a textual description. The robot must then infer and adapt to the remaining tasks based on the diverse objects available in the scene. **Collaboration type 2** presents a higher level of difficulty. In this type, one intention (e.g., morning hygiene) is broken down into 5 tasks, where the human performs free-form motion while interacting with a static object (e.g., human brushes teeth in front of the mirror), and the robot is tasked with offering a desired object from a magic box (e.g., robot offers a toothbrush). For each intention, the robot is provided only a video of the human performing the motion of the first task, without textual descriptions.

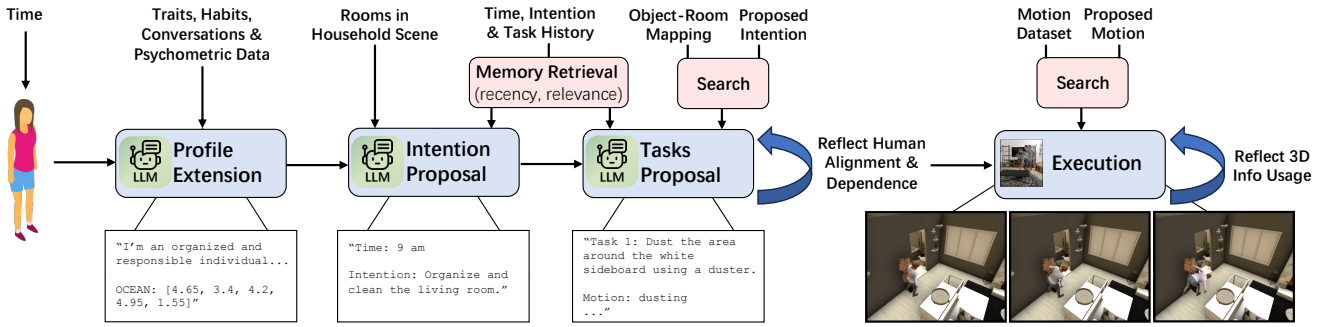

Figure 3. **Human Simulation Pipeline.** We seed the human-LLM with an extended profile. At each time of day, the human proposes an intention and decomposes it into tasks, aligning with profile traits and temporal dependence on intention/task history. LLM inputs are optimized with *Memory Retrieval* 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 the Appendix for detailed prompts.

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, Human 2, and Human 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, Human 2, and Human 3 in the first scene, then repeating this sequence in the second and third scenes (9 days, 3 scenes).

3.2. Simulating Traits-Driven Humans

Modeling Unique Humans. We simulate the human agents using LLMs since our goal is endow human agents with unique personalities and traits that drive their behaviors in interactions with others and the environment [21, 54]. We enable the human-LLMs to generate behaviors aligned with specific personalities and traits by prompting simulated conversations with other virtual humans, fetched from the synthetic human dataset [21]. This process allows the human-LLM to generate an extended summary of the human profile in a detailed paragraph. Additionally, we incorporate psychometric data [16] as an attribute in the human profile, inspired by previous work [37, 49, 56] showing that LLMs have human-like abilities to understand psychological dispositions. Psychometric data will also serve as one of the evaluation metrics, detailed in Section. 4.

Whole-Day Intentions and Tasks. At the core of our framework are human agents that exhibit whole-day intentions and perform tasks featuring temporal dependence and varying distributions. *Temporal Dependence:* Given the current time and 3D envrionment information, the human-LLM proposes intentions throughout the day and decomposes them down into a set of inter-dependent tasks. During intention and task proposals, the human-LLM is also provided with the history of intentions and tasks from previous hours, and is explicitly prompted to consider their interdependence. *Varying Distribution:* Although a human with specific psychological traits and habits follows a general routine, their daily behavior is not strict (i.e., Monday at 9 am for cleaning, Tuesday at 9 am for exercise). Therefore, we reset the intention and task history at the start of each new day, setting a high temperature for the human-LLM to simulate variability in human intentions and tasks.

Expressive Whole-Body Motion. We also simulate our human agents physically using 3D expressive whole-body motions during task executions [19, 25, 31] by chaining a series of motions for each task. For pick-and-place tasks, the human motion sequence consists of walking, reaching and picking, walking again, and then reaching and placing. For tasks involving free-form motion (e.g., sitting on a sofa), during the task proposal stage, the human-LLM describes a free-form human motion that matches each task, using a list of available motions from our human motion dataset as examples. The resulting sequence combines walking with the appropriate free-form human motion.

Optimizing Long-Context Inputs. In our progressive prompting chain, the human-LLM is given substantial information at each stage. Particularly during the task proposal stage, where hundreds of objects may exist in the 3D environment, and our human motion dataset contains thousands of data points. Additionally, as the day progresses, the intention and task history can grow long (e.g., from 9 am to 9 pm, 13 intention sentences and dozens of task descriptions accumulate). Since LLMs struggle with very long context inputs [24, 27, 51], we introduce mechanisms of Search and Memory Retrieval. *Search:* Given a query text and a list of texts, we return the top-K most relevant items based on

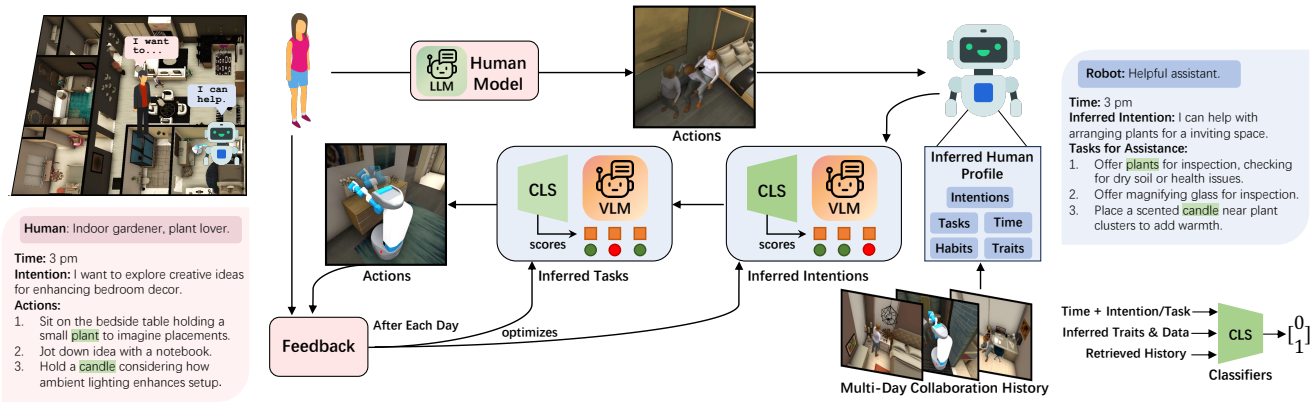

Figure 4. **Our approach of human imitation.** We decouple robot task inference into two stages: first inferring intentions, then identifying specific tasks. By chaining VLM and classifier, the robot gradually filters and selects tasks that best correlate with the human's traits and temporal context. The robot keeps track of a human profile, inferred from collaboration history. This profile, combined with human feedback, optimizes the robot-VLM through prompting and the classifiers through supervised learning. Please see Appendix for more details.

semantic similarity. *Memory Retrieval:* We use recency and relevance scores to retrieve the top-K memories. Recency scores decay over time from the current time with a decay factor λ, while relevance scores are calculated by semantic similarity, similar to the search mechanism. The final retrieval score is the product of the recency and relevance [35].

Self-Corrections. Given the complexity of our progressive prompting process and human-LLM responses, mistakes can occur even in state-of-the-art (SOTA) LLMs. Therefore, during the most complex task proposal stage, we perform two rounds of *Reflexion* [45, 52] to identify and correct errors related to human traits, temporal dependencies, and object use within the 3D environment.

3.3. Building an Assistive Agent

Evaluation Metrics. We assess the performance of the assistive agent by **success rate** (i.e., given the ground truth human intention and a set of tasks inferred by the robot, we ask the simulated human if the intention is fulfilled).

Main Method (Human Imitation). Our approach (Fig. 4) enhances the robot's collaboration with humans by enabling it to mimic human traits and habits, learning correlations between human intentions and tasks at each time of day, and their traits and temporal dependencies.

At each time of day, a human's intentions/tasks can be viewed as meta-intentions/meta-tasks, encompassing a range of possible options. Our solution decouples task inference into two stages: first inferring intentions, then identifying specific tasks. We capture the correct sets by chaining VLM for imagining multiple possible intentions/tasks and classifiers for scoring and filtering. Specifically, given observation (frames uniformly extracted from a video $V = [f_1, f_2, \ldots, f_N]$) of the human's first task execution, the robot-VLM generates an intention superset. For each intention that is positively classified 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.

Baselines. As there is no prior work presenting a solution for our framework, we adapt a combination of standard approaches and modifications of our approach as baselines. **1) Direct prompting:** The robot proposes a single intention based on visual observations and decomposes it into tasks. The robot-VLM is optimized solely through prompting retrieved history of human intentions and tasks. **2) Direct finetuning:** The robot brain is finetuned to directly output a single intention and decompose it down into tasks. **3) GT_Intention** (replacing imagination modules): We provide the robot with the ground truth human intention and allow it to decompose the intention into a single set of tasks. **4) Random** (removing classifiers): We eliminate the intention and task classifiers, and instead consider all multiple proposed intentions and tasks by the robot as correct without further validation. **4) 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.

4. Experiments

We conduct experiments to answer the following questions: 1) How accurately does our simulated human model reflect real human behavior? 2) Does the robot's assistance becomes more personalized after days of collaboration in various settings? 3) How effective is each module in our framework?

4.1. Dataset

Environment and Scene. We use *Habitat 3.0* [41] as the robot simulation platform and *Habitat HSSD* [23] as the 3D environment, which features diverse scenes in terms of style and size, and contains a large number of objects (18,656). Since the original Habitat 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) dynamic. Additionally, we randomly sample and introduce 20 dynamic objects from the YCB dataset [7] for each scene. 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 both the human and Fetch robot [13] to navigate.

Human. For modeling unique humans, we use the *SPC: Synthetic-Persona-Chat Dataset* [21], a fully synthetic dataset that includes hundreds of short user profiles along with their conversations. Since SPC lacks psychometric data, we derive Big-5 OCEAN scores by prompting the LLM to 1) directly infer the scores [37] and 2) complete the Big-5 personality test [16, 34] 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. Note that for the human modeling processes, we use the open-source LLM, Llama-3.1-70B [11], to enhance interpretability. Finally, we use Motion-X [25] and AMASS [30] as the human motion dataset. We generate 10 human profiles.

4.2. Implementation Details

Human Simulation. We use GPT-4o [1] as the human-LLM, with the temperature set to 0.7. For search and memory retrieval, we utilize the MiniLM-L6-v2 model [48]. The decay factor λ is set to 0.95, and we retrieve the top 5 tasks during memory retrieval.

Assistive Agent. Task videos from Habitat 3.0 [41] are resized to 1024×768, with 3 frames as input. GPT-4o [1] is employed as the robot-VLM. For training intention and task classifiers, we finetune Mistral-7B-Instruct-v0.2 [22] in an instructional finetuning manner to output binary yes/no answers using LoRA [20]. The rank is set to 8, dropout to 0.2, and alpha to 16, targeting the q, v, k, and o projections. We train for 3 epochs and use the AdamW optimizer [28] with a base learning rate of 1e-5, weight decay of 0.01, and a per-device batch size of 1. All training and inference are run on 3 NVIDIA A10 GPUs, each with 24GB RAM.

Table 1. **Evaluation of 1)** Human classification, **2)** Human traits-psychometrics coherence, **3)** temporal dependence, and **4)** real-human experiments.

Classification (Acc ↑)		**Coherence** (R ↑)		**Temporal Dependence**		**User Study** (Acc ↑)	
intention	task	aligned	mismatched	Acc ↑	F1 ↑	MCQ	Matching
0.995	0.830	0.319	-0.482	0.789	0.790	0.764	0.712

4.3. Human Simulation

Distinct Simulated Humans. We explore whether simulated humans with varying human traits and Big-5 personality scores exhibit distinct features that can be identified by machines. For each of the 10 simulate humans, we place them in 5 different scenes, living for 20 days in each scene. For each human, we aggregate their proposed intentions and tasks over a single day into one data point We then finetune two 10-way text classifiers (BERT-large-uncased [9])—one for intentions (10 epochs) and one for tasks (20 epochs)—to distinguish between individuals, using an 0.8:0.2 split for training and testing with a start learning rate of 5e-6. Testing is performed on an unseen scene. From Table. 1, classifying human tasks proves more challenging than intentions, as intentions are more closely aligned with human traits, while tasks (e.g., drinking water) can correspond to multiple intentions, such as leisure or exercise.

Human Traits and Psychometrics Coherence. In our main approach of human imitation, the robot-VLM infers the Big-5 personality scores of the collaborating human throughout the day, based on the collaboration history of human intentions and tasks. We use the inferred scores at the end of the collaboration as the final scores. To evaluate the coherence between these inferred scores and the ground truth, we calculate the Pearson correlation, following Peters *et al.* [37] and Azucar *et al.* [4]. Additionally, we intentionally shift the order of the ground truth and inferred scores by one to create a mismatch and compute the averaged correlation. This comparison assesses the coherence between human traits and the inferred Big-5 personality scores. The significant decrease in mismatched pairs compared to aligned pairs in Table. 1 indicates the coherence between human traits and psychometric data in our human model, and the LLM's ability to interpret human psychology from behavior.

Temporal Dependence in Human Behavior. We investigate the dependency of current-hour intentions on those from previous hours by conducting a next-sentence prediction task. Given three intentions from prior hours (e.g., 9-11 am), we use the intention at the current hour (e.g., 12 pm) as the positive example and intentions from other times as negatives. We evaluate this on 10 simulated humans, each placed in 5 different scenes and living for 20 days in each scene. We train BERT-large-uncased [9] for 20 epochs with start learning rate 5e-6. Results in Table. 1 demonstrate the observed temporal dependence.

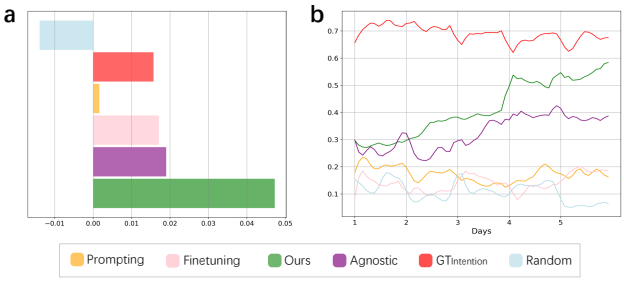

Figure 5. **Evaluation of changes in robot success rate:** a) within a single day and b) across multiple days. See Appendix for a detailed breakdown of the results.

Real-Human Experiments. We conduct two user studies to assess: 1) Whether the same simulated humans across different days and scenes can be identified by real humans, and 2) Whether simulated humans with varying traits and Big-5 scores can be distinguished by real humans. For the first study, we randomly sample full-day intentions and tasks for 10 simulated humans across 2 days and 2 scenes (4 samples per human). We then construct 10 multiple-choice questions, where each question presents the human profile and an example of a full-day human behavior, and the three options include one correct full-day behavior of the same human and two incorrect ones from other humans. For the second study, we randomly sample full-day intentions and tasks from 10 simulated humans, each with distinct traits and Big-5 scores. Participants are then asked to match human trait descriptions with the corresponding full-day intentions and actions. For both studies, we recruit 25 participants. Results in Table. 1 show that, surprisingly, real humans are better at identifying the same simulated human across instances than distinguishing between different simulated humans with distinct traits. We believe this may be because, in our user study, the matching problem is inherently more challenging than multiple-choice questions.

4.4. Human-Robot Collaboration

Setup. We follow the setting, evaluation metrics, and baselines outlined in Section. 3. In Setting 1, we evaluate 5 humans across 2 scenes. For Setting 2, we evaluate 10 humans. Setting 3 is evaluated across 3 distinct scenes, and Setting 4 is evaluated across 9 humans.

Results. We are interested in two aspects: 1) Within-day improvement—does the robot's collaboration success rate increase as the day progresses? This evaluates if the robot is able to learn the temporal dependencies between human intentions and actions over the course of a day. 2) Across-day improvement—does the robot's collaboration become more successful and personalized across multiple days? This assesses if the robot can learn human traits and habits, utilizing feedback provided at the end of each day. From Fig. 5 (a), our main method has the highest within day improvement,

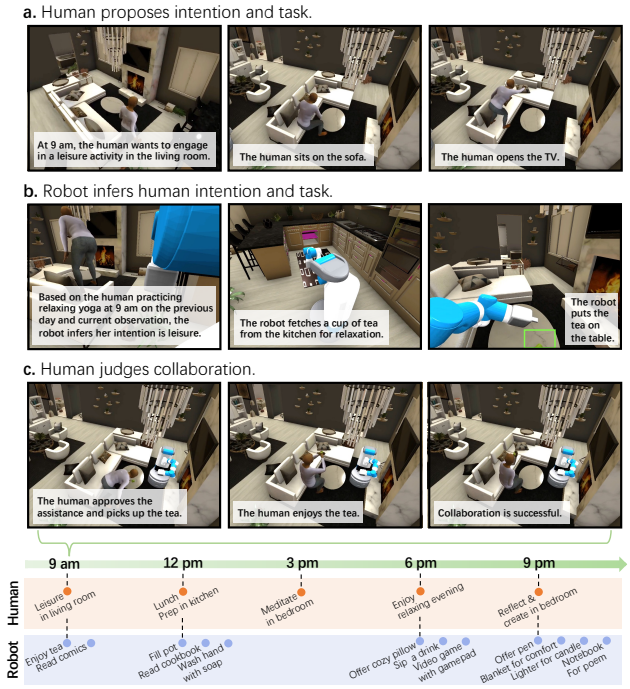

Figure 6. **Qualitative examples of successful human-robot collaboration within one day.** The red row displays human intentions, while the blue row shows the robot's inferred tasks for assistance.

while gt_intention, finetuning, and human & context agnostic show some degrees of improvement. In contrast, random and prompting show little to no improvement, or even a decrease. We hypothesize that this is because the method of random and prompting methods do not benefit from learning human intentions, which are fundamentally more correlated with human traits and temporal context. This observation is also supported by our human classification experiment in Section. 4.3. From Fig. 5 (b), gt_intention and random achieve the highest and lowest success rates, as expected. Our method shows the greatest trend of improvement across days, second only to gt_intention. The minimal improvement seen in prompting and finetuning highlights the challenge of varying human behavior across days, as these methods tend to establish a one-to-one mapping between time and human intentions/tasks. Prompting relies heavily on previous collaboration history as a strong indicator, while finetuning tends to prioritize the highest-probability training data, limiting adaptability to varying human behaviors.

4.5. Qualitative Results

We showcase how the robot's collaboration with the human improves within a single day by correctly inferring more tasks for assistance, along with a visualization of the HRC at a specific time in Fig. 6. Additionally, We present examples of full-day intentions and tasks proposed by a human with specific human traits and psychometric data in Fig. 7.

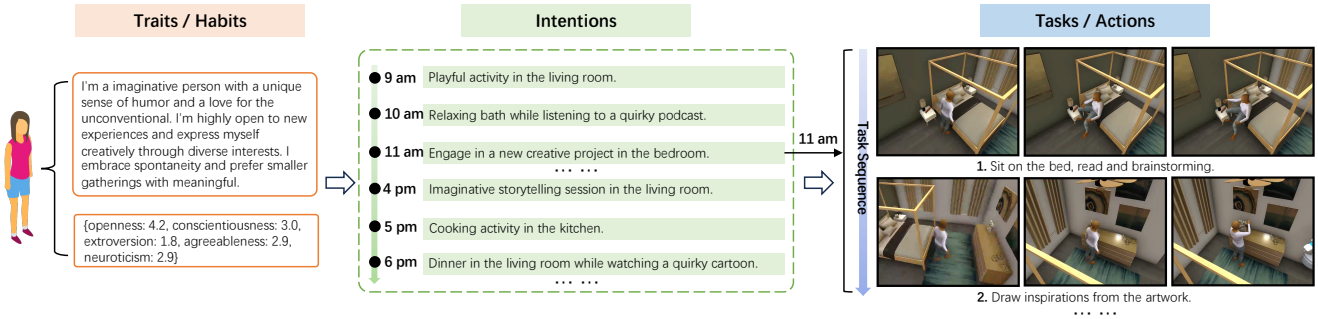

Figure 7. **Qualitative examples of full-day intentions and tasks** proposed by a human with specific human traits and psychometric data.

Table 2. **Ablation study on human simulation:** Evaluation of the effects of removing profile extension via human classification and the impact of single-shot intention proposal on temporal dependence through next-intention prediction. Accuracy change is reported on a scale of 0 to 1.

Removing Profile Extension (ΔAcc $\downarrow$)		**All-Day Intention Proposal**	
intention	task	ΔAcc $\downarrow$	ΔF1 $\downarrow$
-0.045	-0.030	-0.038	-0.050

Table 3. **Ablation study on building an assistive agent.** We report the success rate $\uparrow$ averaged over the last day of collaboration.

	Setting 1	**Setting 2**	**Setting 3**	**Setting 4**
No Traits	0.513	0.482	0.163	0.188
No Context	0.481	0.438	0.471	0.452
Changing Backbone	0.448	0.412	0.376	0.345
Ours (main)	0.552	0.541	0.513	0.488

4.6. Ablation Studies

Removing Human Profile Extension. For modeling unique humans, we explore whether our approach creates the most distinct simulated human profiles. We exclude simulated conversations with other virtual humans and skip generating an extended summary of the human profile. Instead, we prompt only with the short human traits paragraph from the dataset. We assess the effects by finetuning and calculating the accuracy of the human intention and task classifiers, as detailed in Section 4.3. Results in Table. 2 proves the effectiveness of our profile extension module.

Single-Shot Human Intention Proposal. We investigate whether our pipeline design for the human model maximizes temporal dependence between whole-day intentions and tasks. Rather than having the simulated human propose intentions by explicitly considering previous intentions and task history hour by hour, and using Reflexion, we remove these components, allowing the human to propose whole-day intentions all at once. We evaluate the effects by next-intention sentence prediction, as detailed in Section 4.3. Results in Table. 2 shows the effectiveness of our approach.

Removing Human Traits Inference. We examine the importance of learning human traits. In our main method of human imitation, the robot no longer infers human traits or Big-5 personality scores based on the history of human intentions and tasks. In this setup, the classifiers do not learn the correlation between human traits and intentions/tasks. The results in Table. 3 show a significant drop in success for settings 3 and 4, which involve multiple humans. This decrease occurs because the robot struggles to distinguish between different humans.

Removing Temporal Context Learning. We examine the importance of learning temporal dependence between human intentions and tasks. In our main method of human imitation, the classifiers no longer learn the temporal dependence between previous intentions/tasks and the current intention/task, removing time-based context from decision-making. Results in Table. 3 suggest that learning context information are useful across all settings.

Changing the Robot Brain Backbone. We replace the robot brain's VLM from GPT-4o [1] to Llama-3.2-11B [11], and reduce the robot's observation to a single input image. Using different VLMs for the human and robot brains reduces alignment, allowing us to test the robustness of our approach. Despite the smaller VLM size, the robot's success rate still reaches a reasonable level, as shown in Table. 3.

5. Conclusion

In this work, we introduce a framework for continual open-ended human-robot collaboration. We propose a model to generate long-term human behaviors driven by personality traits, and a method to assist humans under such setting 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.

References

[1] Open AI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. 6, 8

[2] Stefano V. Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, pages 66–95, 2018. 2

[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. 2

[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. 6

[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. 2

[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. 2

[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. 6

[8] 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. 2

[9] 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. 6

[10] 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. 2

[11] 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. 6, 8

[12] Alan Fern, Sriraam Natarajan, Kshitij Judah, and Prasad Tadepalli. A decision-theoretic model of assistance. *Journal of Artificial Intelligence Research*, 50:71–104, 2014. 2

[13] Fetch Robotics. Fetch, 2020. 6

[14] 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. Feigelis, 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. 2

[15] Théophile Gervet, Soumith Chintala, Dhruv Batra, Jitendra Malik, and Devendra Singh Chaplot. Navigating to objects in the real world. *Science Robotics*, 2023. 2

[16] L R Goldberg. An alternative "description of personality": the big-five factor structure. *Journal of Personality and Social Psychology*, pages 1216–1229, 1990. 4, 6

[17] Michael A. Goodrich and Alan C. Schultz. Human-robot interaction: A survey. *Foundations and Trends in Human-Computer Interaction*, pages 203–275, 2007. 2

[18] 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. 2

[19] 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. 4

[20] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 6

[21] 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. 2, 4, 6

[22] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas,

Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 6

[23] 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. 6

[24] Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*, 2024. 4

[25] 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. 4, 6

[26] 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. 2

[27] 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. 4

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[29] 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. 2

[30] 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. 6

[31] 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. 4

[32] 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. 2

[33] 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. 2

[34] OpenPsychometrics. The big five personality test, n.d. 6

[35] 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. 2, 5

[36] 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. 1

[37] Heinrich Peters and Sandra C. Matz. Large language models can infer psychological dispositions of social media users. *arXiv preprint arXiv:2309.08631*, 2023. 4, 6

[38] 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. 2

[39] 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. 1, 2, 3

[40] 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. 2

[41] 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. 1, 2, 6

[42] 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. 2

[43] 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. 2

[44] 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. 2

[45] 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. 5

[46] 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. 2

[47] 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. 1

[48] 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. 6

[49] 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. 4

[50] 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. 2

[51] 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. 4

[52] Mert Yüksekgönül, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. Textgrad: Automatic "differentiation" via text. *arXiv preprint arXiv:2406.07496*, 2024. 5

[53] 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. 1

[54] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Annual Meeting of the Association for Computational Linguistics*, page 22042213, 2018. 4

[55] 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. 2

[56] 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. 4