COOPERA: Continual Open-Ended Human-Robot Assistance

Chenyang Ma[1], Kai Lu[1], Ruta Desai, Xavier Puig, Andrew Markham[1], Niki Trigoni[1]

[1] University of Oxford

NeurIPS 2025 (Spotlight)

Motivation

- Robots that can assist humans in their daily lives by adapting to their preferences and habits.
- Need to free/assist human: consider the scenario of a robot in a house with no explicit, specific goals.
- Need to understand human intentions, reason in long-term, adapt to human behavior.

- Pre-defined tasks
- Episodic setting
- Static environment

Vs.

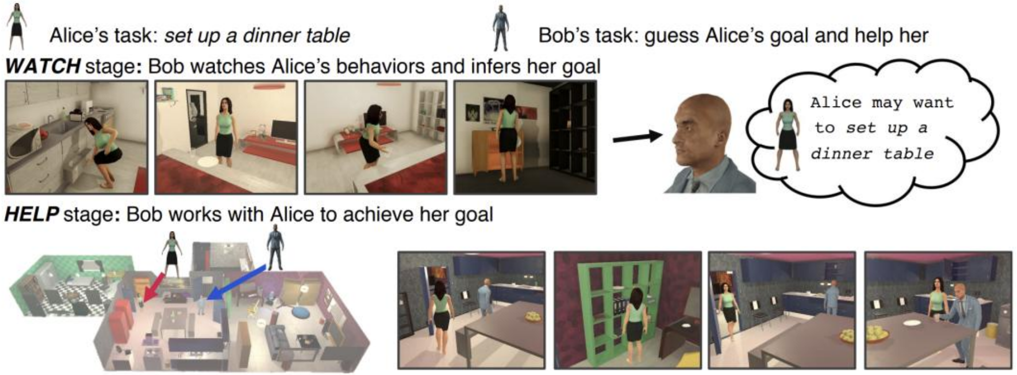
- Only high-level, abstract goals
- Long-term
- Dynamic environment (i.e., interact with human)

I can help her to get a hot coffee.

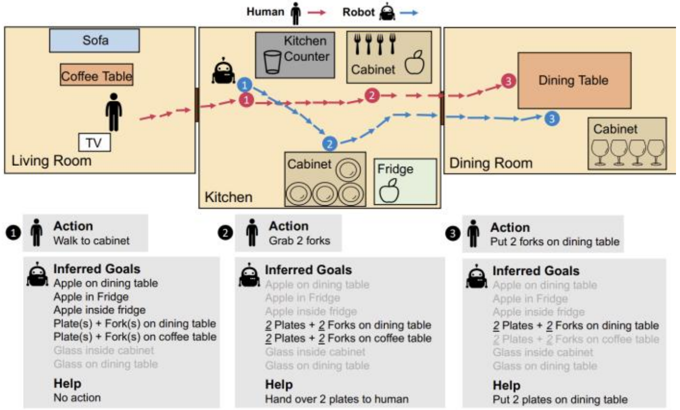
I should open the drawer in advance.

Previous Works on Human-Robot Collaboration

- Predefined human intentions.
- Close-set tasks.
- Episodic settings.

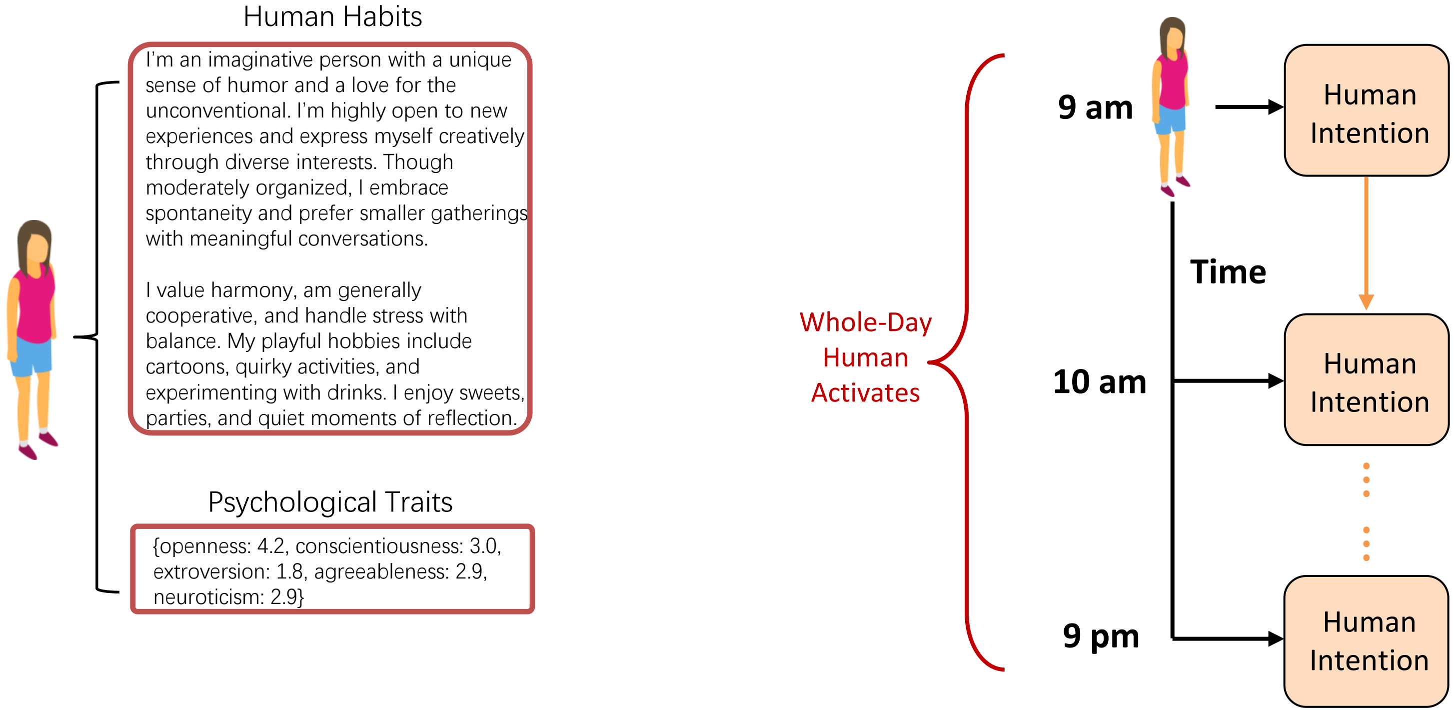


Watch-And-Help (2021)



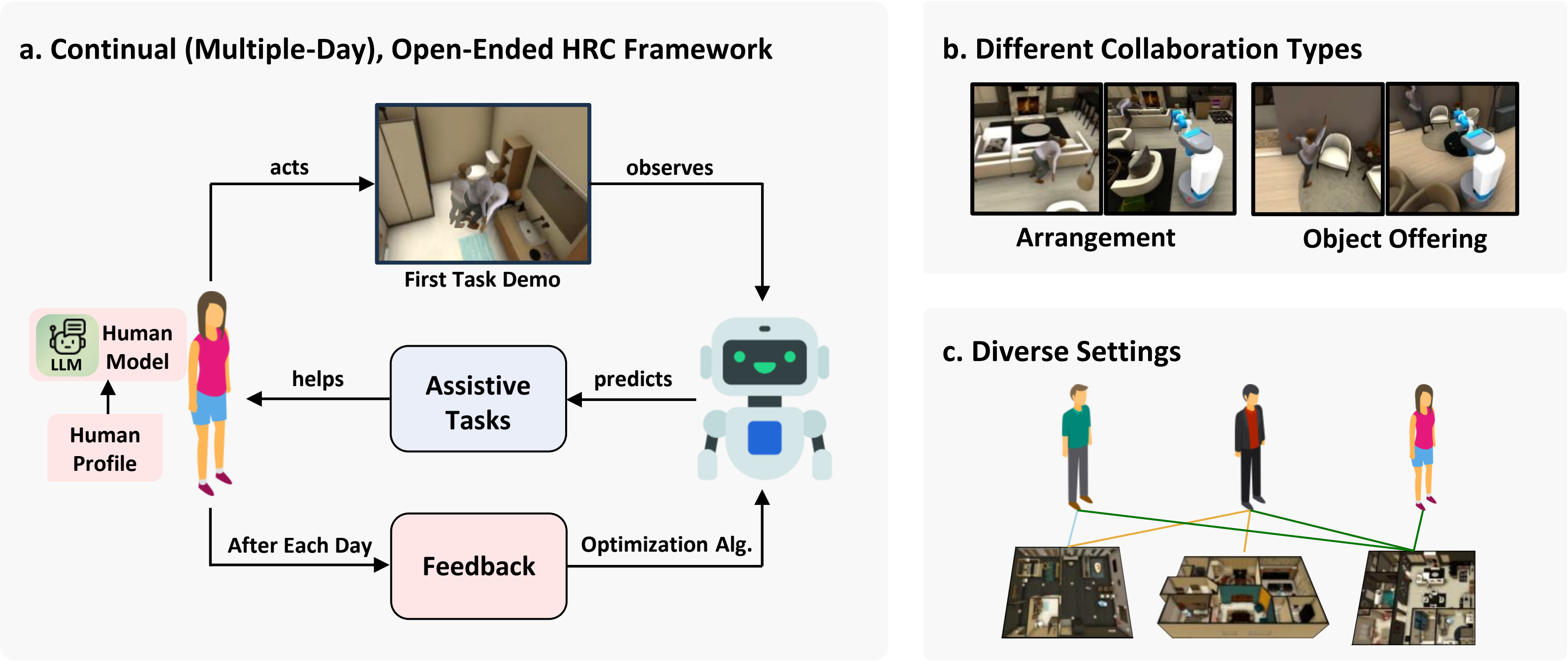
NOPA (2023)

Motivation: Towards Continual and Open-Ended Human-Robot Collaboration



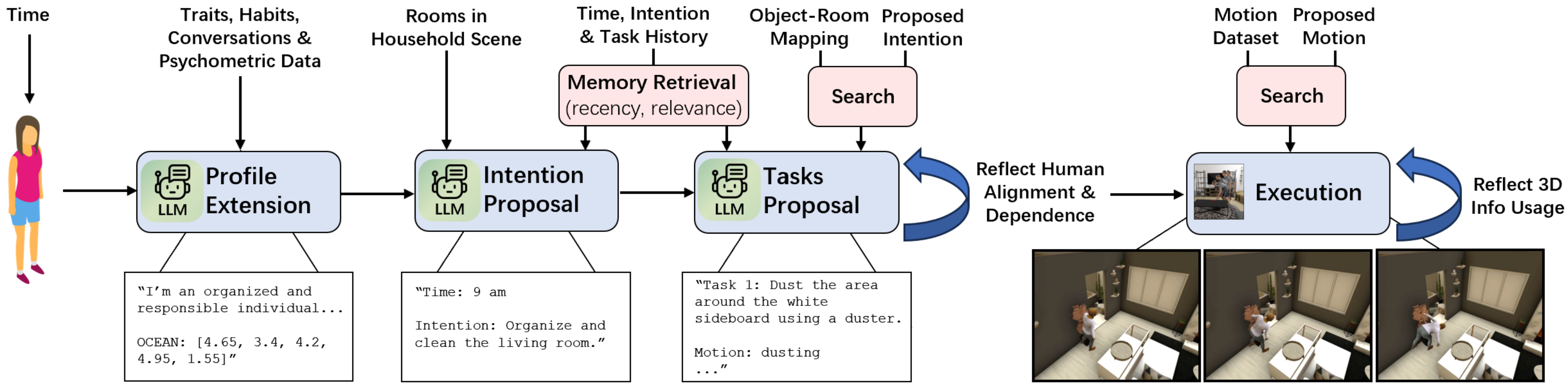
- **Traits-Driven Human Modeling.**
  - Humans interact in the environment driven by intentions that unfold over one day (i.e., whole-day).
  - These goals and intentions should be open-ended and contextualized.
  - Driven by psychological traits and habits.

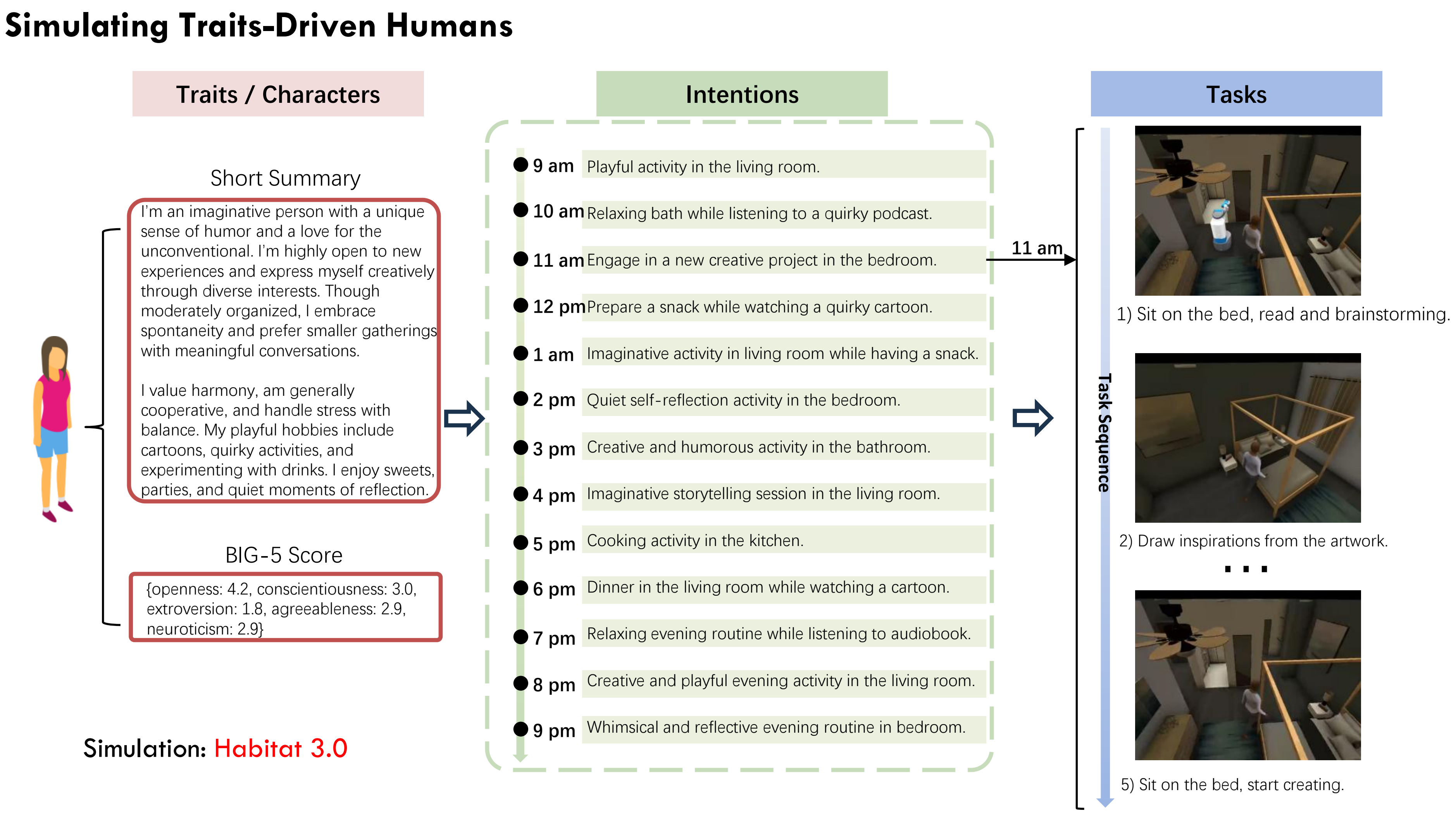
Motivation: Towards Continual and Open-Ended Human-Robot Collaboration



- **Continual Collaboration**
  - Day's beginning: the robot observes and collaborates with the human, assisting in inferred tasks.
  - Day's end: the human communicates with the robot and provides feedback.
  - This enables collaboration over multiple days, and improves HRC success rate for subsequent days.

Simulating Traits-Driven Humans





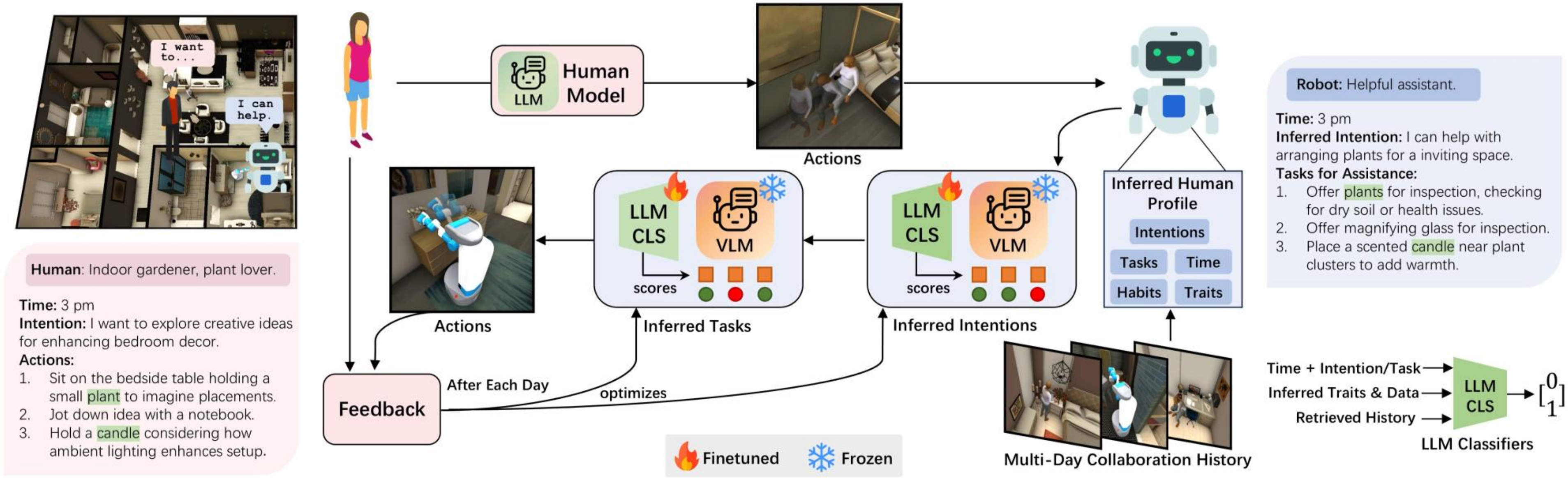
Simulating Traits-Driven Humans
Traits / Characters
Short Summary
I'm an imaginative person with a unique sense of humor and a love for the unconventional. I'm highly open to new experiences and express myself creatively through diverse interests. Though moderately organized, I embrace spontaneity and prefer smaller gatherings with meaningful conversations.
I value harmony, am generally cooperative, and handle stress with balance. My playful hobbies include cartoons, quirky activities, and experimenting with drinks. I enjoy sweets, parties, and quiet moments of reflection.
BIG-5 Score
{openness: 4.2, conscientiousness: 3.0, extroversion: 1.8, agreeableness: 2.9, neuroticism: 2.9}
Simulation: Habitat 3.0
Intentions
9 am Playful activity in the living room.
10 am Relaxing bath while listening to a quirky podcast.
11 am Engage in a new creative project in the bedroom.
12 pm Prepare a snack while watching a quirky cartoon.
1 am Imaginative activity in living room while having a snack.
2 pm Quiet self-reflection activity in the bedroom.
3 pm Creative and humorous activity in the bathroom.
4 pm Imaginative storytelling session in the living room.
5 pm Cooking activity in the kitchen.
6 pm Dinner in the living room while watching a cartoon.
7 pm Relaxing evening routine while listening to audiobook.
8 pm Creative and playful evening activity in the living room.
9 pm Whimsical and reflective evening routine in bedroom.
11 am
Tasks
Task Sequence
1) Sit on the bed, read and brainstorming.
2) Draw inspirations from the artwork.
...
5) Sit on the bed, start creating.

Challenges: Instantiating COOPERA with an Assistive Agent

- 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.
  - They need to adapt to each person's preferences and traits.

- There is one-to-many mapping between the time and human behavior, even for the same human.
  - On Monday 9 am, the human may take a shower. On Tuesday 9 am, the human may eat breakfast.

Instantiating COOPERA with an Assistive Agent



Experiments

- How accurately does our simulated human model reflect real human behavior?
- Does the robot’s assistance become more personalized after days of collaboration in various settings compared to other methods in our introduced benchmark?
- Can COOPERA be applied and tested in real world?
- How effective is each module in our framework?

Analysis of Human Simulation

- Distinct Simulated Humans: text classification.
- Diverse Simulated Humans: averaged per-trait standard deviation.
- Human Traits and Psychometrics Coherence: robot VLM inference vs. ground truth.
- Temporal Dependence in Human Behavior: next intention prediction.
- Realisticness: user study.
- Alignment with Real-Human Behavior: paragraph semantic similarity.

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

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

	Generic	**Mismatched**	**Main**
SBERT ↑	0.554	0.523	0.810
OpenAI Emb. ↑	0.537	0.543	0.772

Analysis of Continual, Open-Ended HRC

- Within day improvements.
- Across days improvements.
- Out-of-domain generalization.

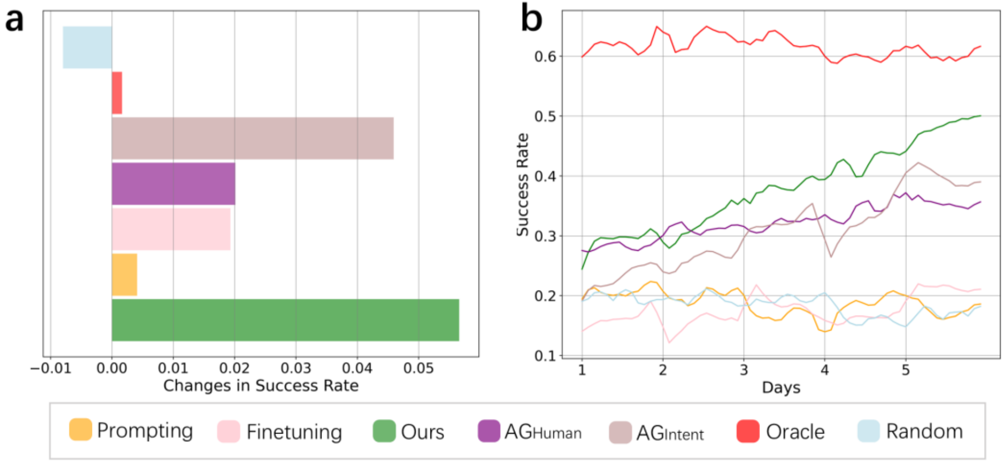


Table 3: **Generalization performance.** We report the average success rate (predicate-based).

	Baseline	Finetuned
Scene	0.269	0.465
Human	0.258	0.343

Analysis of Real-World Applicability

- Human Verification.
- Collaborating with Offline Real Humans.
- Human-in-the-Loop.

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

Ablation Studies

Human Simulation

- Removing human profile extension.
- Single-shot human intention proposal.

Assistive Agent

- Removing human traits inference.
- Removing temporal context learning.
- Changing the robot brain backbone.

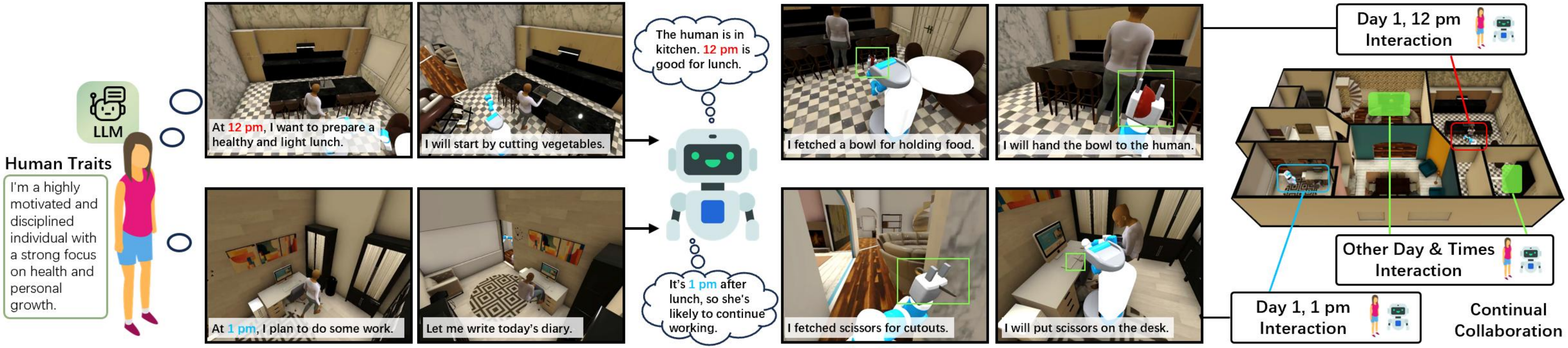
Table 5: **Ablation study on human simulation.** The effects of removing profile extension on human classification and using single-shot intention proposal on temporal dependence.

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

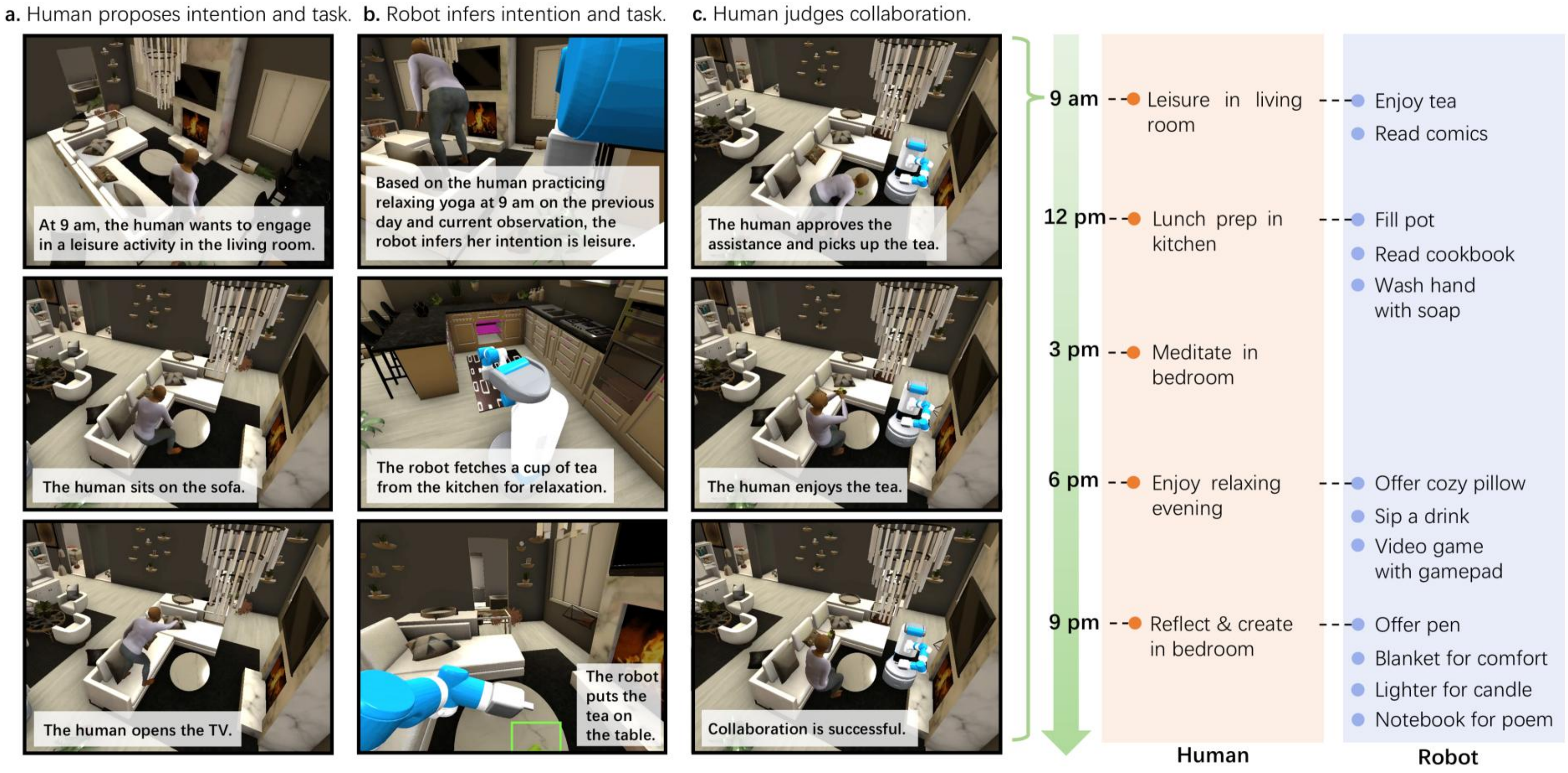
Table 6: **Ablation study on assistive agent:** success rate (predicate-based) ↑ averaged over the last day.

	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

Qualitative Results



Qualitative Results



Thanks for Listening!

Q&A