2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Mixed-Initiative Dialog for Human-Robot Collaborative Manipulation

Albert Yu[1,†], Chengshu Li[2], Luca Macesanu[3], Arnav Balaji[1], Ruchira Ray[4], Raymond Mooney[1], Roberto Martín-Martín[1]

***Abstract*— Effective robotic systems for long-horizon human-robot collaboration must adapt to a wide range of human partners, whose physical behavior, willingness to assist, and understanding of the robot's capabilities may change over time. This demands a tightly coupled communication loop that grants both agents the flexibility to propose, accept, or decline requests as they coordinate toward completing the task effectively. We propose MICoBot, a system that enables the human and robot, both using natural language, to take initiative in formulating, accepting, or rejecting proposals on who can best complete different steps of a task. To handle diverse, task-directed dialog, and find successful collaborative strategies that minimize human effort, MICoBot makes decisions at three levels: (1) a meta-planner considers human dialog to formulate and code a high-level collaboration strategy, (2) a planner optimally allocates the remaining steps to either agent based on the robot's capabilities (measured by a simulation-pretrained affordance model) and the estimated human's willingness to help, and (3) an action executor decides the low-level actions to perform or words to say to the human. In physical robot trials with 18 unique human participants, MICoBot significantly improves task success and user experience over a pure LLM baseline and standard agent allocation models. See additional videos and materials at our project site. [1]**

I. Introduction

Imagine preparing for a dinner party with a friend. Your friend might excel at mixing drinks while you focus on cooking the main dish. You are also better at decorating, while both of you reluctantly negotiate over less desirable tasks like cleaning.

Now, imagine a helper robot in place of the friend. Current robots are not fully autonomous for many household tasks, but they offer broad capabilities with varying levels of reliability that can be leveraged through collaboration with humans. To be an effective partner, such a robot must communicate in physically grounded natural language, decide when to take initiative or defer to the human, negotiate task allocation based on strengths and preferences, and adapt to changing contexts. These ingredients are essential not only for collaborative household robots, but also for coding assistants, chatbots, and AI agents more broadly.

Long-horizon tasks, such as preparing for a party, require dynamic, bidirectional collaboration across control, initiative, and communication. In particular, the ability to both take initiative and yield control is central to effective human–AI teamwork. However, current AI systems (e.g., chatbots) typically rely on one-directional, human-initiated interactions [1, 2], while prior human–robot interaction (HRI) approaches often assume fixed collaboration plans and full human compliance [3]. Such assumptions limit flexibility and fail to account for the diverse preferences, capabilities, and strengths of different human partners. We argue that effective human–robot collaboration requires a paradigm shift toward mixed-initiative dialog as the communicative medium, enabling both agents to initiate, negotiate, and respond to proposals in natural language.

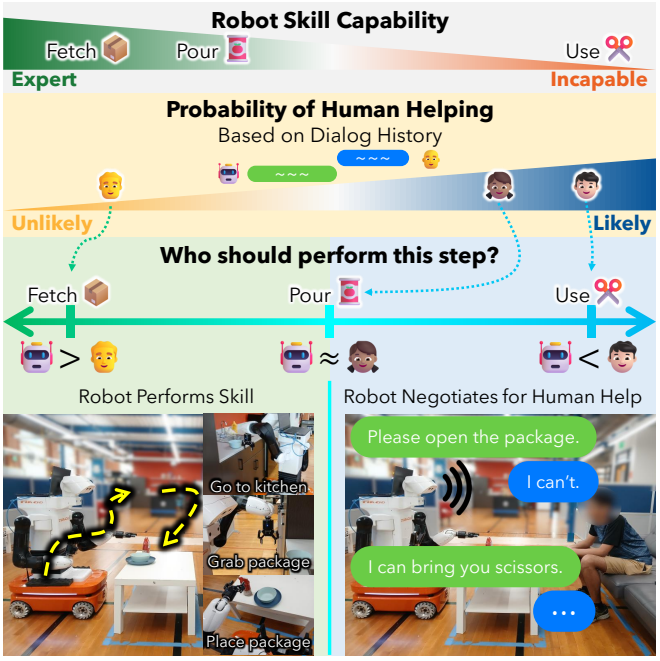

Fig. 1: We present MICoBot, a system for human-robot collaboration where both agents can initiate and carry out physical and verbal actions. MICoBot uses both the robot's capability and the likelihood of human helping (inferred from previous dialog history) to determine whether the robot is better suited than the human to perform the skill. If it is, it attempts the skill itself. If not, it negotiates for human help.

To enable this paradigm shift, we introduce MICoBot (Mixed-Initiative Collaborative roBot), the first system that supports mixed-initiative dialog for seamless human–robot collaboration in the physical world. MICoBot allocates task steps to the most suitable agent (see Fig. 1) in a way that maximizes overall success, minimizes human effort, and respects human-initiated requests. It achieves this by engaging in mixed-initiative dialog and negotiation to decide step allocation (see Fig. 2), while coordinating the physical and verbal actions required to execute the plan.

[1]UT Austin, [2]Stanford University, [3]NYU, [4]University of Edinburgh, [†]albertyu@utexas.edu

[1]https://robin-lab.cs.utexas.edu/MicoBot/

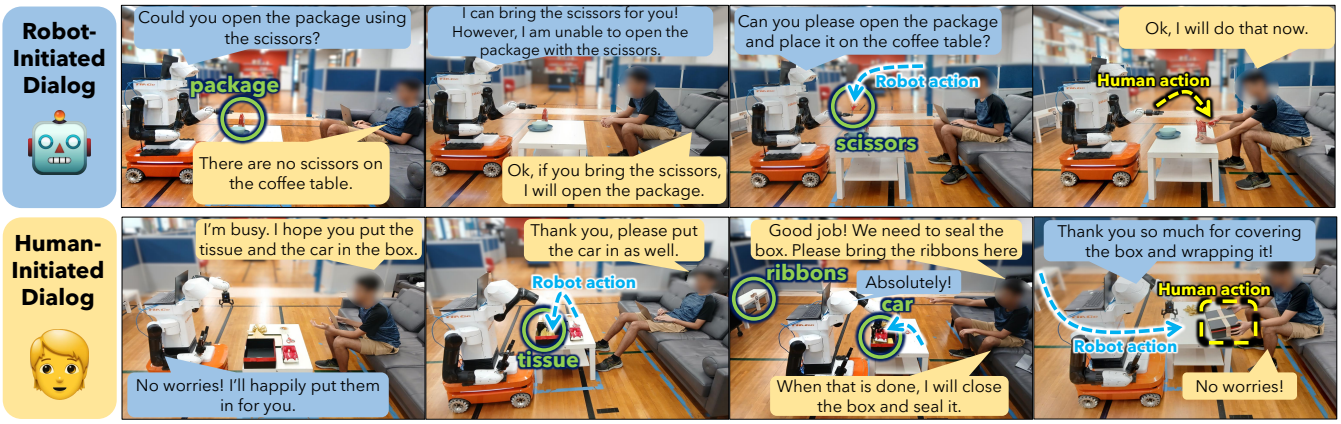

Fig. 2: MICoBot supports *both* robot-initiated (top row) *and* human-initiated (bottom row) task-directed speech2speech dialog, where both agents discuss who is best suited to perform steps in a long-horizon task. These are real dialog and physical interactions from our user studies (see our website[1]).

To realize this objective across diverse humans and long-horizon tasks, MICoBot optimizes decision-making at three levels. First, a meta-planner determines the high-level collaboration strategy, integrates human-specified preferences, and generates adaptive code for robot actions (verbal or physical). Second, a planner executes this code, selecting the optimal collaboration approach based on the environment state, a simulation-trained affordance model of robot capabilities, and a dynamic estimate of human helpfulness derived from prior interactions. Finally, an action executor carries out the next step of the plan, whether it involves manipulation, dialog initiation, or dialog response.

We validate MICoBot through extensive evaluation in both simulation and the real world. In simulation, we test against LLM-simulated humans with varying helpfulness and responsiveness; in real-world experiments, 18 participants collaborated with a TIAGo mobile manipulator on three household tasks. Our approach improves success rate by **50%** compared to a pure LLM baseline and is preferred by over **75%** of participants.

In summary, our contributions are:

- **A new problem setting** that integrates mixed-initiative natural language dialog with mixed-initiative human–robot interaction.
- **A novel optimization framework** for task allocation that balances human and robot effort with success through a unified metric.
- **A collaborative simulation environment** built on MiniBehavior [4], featuring LLM-controlled virtual humans, with an interactive demo on our project site.[1]
- **A hierarchical robotic system, MICoBot**, that enables mixed-initiative speech-to-speech human–robot collaboration and flexibly adapts to diverse real human collaborators in physically grounded, long-horizon tasks.

II. Related Work

Mixed-initiative dialog [5–7] refers to communication with freeflowing questions and answers from both parties. In the NLP field, the dominant chatbot paradigm adopted by large language models (LLMs) largely eschews mixed-initiative interaction: humans pose substantive questions, and the chatbot primarily responds to fulfill these requests [1, 2]. Recent work has sought to make dialog systems more goal-directed and proactive by incorporating mixed-initiative strategies in tasks such as creating documents [8], persuading users to donate to charity, enhancing users' emotional well-being [9–12], clarifying ambiguous human requests [13–15], or as part of an active-learning framework [16]. However, none of these systems addressed mixed-initiative dialog in grounded, real-world collaborative scenarios involving physical manipulation tasks.

In the human-robot interaction (HRI) field, researchers have developed **human-robot collaboration systems** that interact through language but are restricted to **single-initiative dialog**. Some of these systems integrate LLMs as task planners or delegators [17–19] for tasks like real-world cooking [17] and object sorting [18]. Other systems implement a leader-follower paradigm in simulated worlds, where the leader issues natural language instructions for the follower to execute [20–23]. Single-initiative HRI systems can ask humans for clarification [24] or assistance [25–27], or inform humans of their observations [28–30]. However, by supporting only single-initiative dialog, these systems lack the capacity to adapt to the evolving nature of the human, robot, and environment—limiting their capacity to find the optimal division of labor that respects user preferences [18].

Some works in HRI have explored **mixed-initiative collaborative systems without dialog**, only with physical actions [31–36]. In particular, [37] studied separate regimes of agent initiative (human-initiative, requesting help, or robot-initiative, proactively helping), but failed to support a natural human-robot dialog. By focusing solely on physical actions, these prior works overlook the critical role of communication in effective collaboration, thereby limiting the flexibility of the human-robot team. With MICoBot, we enable both agents to take initiative—through both physical and verbal actions—via task-grounded dialog.

Several prior works in robotics and planning have studied

the problem of **human-robot optimal task allocation**, typically optimizing the time to perform a task or minimizing idle agents, posing the problem as a scheduling optimization [38, 39]. Others have prioritized different objectives, such as safety [40] through the formulation of a constrained optimization problem [41]. While these solutions may achieve shorter execution times, they assume a priori known capabilities and availability of all agents, including both robots and humans. In contrast, MICoBot can adapt to the specific human's willingness to help by estimating its availability based on previous dialog.

III. Problem Setting: Task Collaboration with Mixed-Initiative Dialog

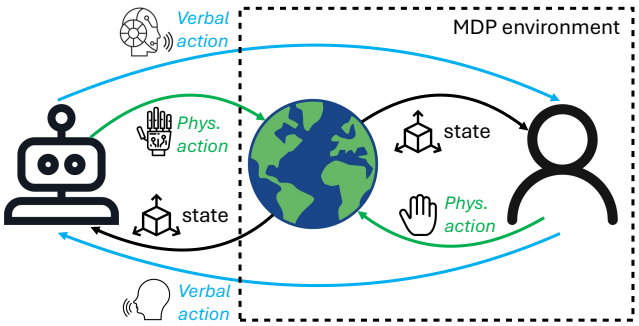

Fig. 3: Proposed MDP for Mixed-Initiative Collaboration.

MDP Formulation. We study how human-robot collaborative manipulation can be facilitated through mixed-initiative dialog. We assume that both agents can observe the state of the world, $s \in \mathcal{S}$, and perform actions, $a \in \mathcal{A} = \mathcal{A}_p \cup \mathcal{A}_v$, comprised of a physical action space, $\mathcal{A}_p$ (e.g., move objects, open them, etc.), that directly affect the physical state of the environment s, and a free-form, natural language verbal action space, $\mathcal{A}_v$, which is directly observed by the other agent but does not change the physical state. We model the problem as a Markov Decision Process (MDP) from the robot's point of view (see Fig. 3), where on each environment step, the robot performs some action, $a_R \in \mathcal{A}_{p,R} \cup \mathcal{A}_{v,R}$ and receives an observation $o = [I, a_{v,H}, s_{proprio}]$ consisting of an RGB-D image I, an optional verbal action from the human partner $a_{v,H}$, and the robot's proprioceptive state $s_{proprio}$. Within each environment step, the human may perform a series of actions, $a_H \in \mathcal{A}_{p,H} \cup \mathcal{A}_{v,H}$, in its own physical and verbal action space after perceiving the world state and robot's previous dialog, $a_{v,R}$.

Physical and Verbal Action Spaces. The physical and verbal action spaces, $\mathcal{A}_p$ and $\mathcal{A}_v$, are shared between both agents. Each element of these action spaces is a parameterized action primitive represented by the pair, $a_{p/v} = (\omega_{p/v}, \theta_{p/v})$. ω_p is the type of the physical action primitive (`open`, `pick-and-place`, etc.) and θ_p are the corresponding parameters (e.g., what object to open or pick and where to place it). We assume that humans are fully competent in executing all steps of a collaborative household manipulation task but may be unwilling or unavailable to perform some or all required actions. Their behavior can range from indifferent (never acting) to overly proactive (completing the entire task without robot involvement).

In contrast, robots often have limited manipulation capabilities and may be unable to execute more complex actions, in which case it uses verbal actions to communicate with the human. ω_v is the type of the verbal action primitive (`ask_human_for_help`, `respond_to_human`, etc.), and θ_v are the corresponding parameters defining the context of the verbal primitive (e.g., what step the robot needs help on). While the types of verbal actions are limited, each generates freeform and open-vocabulary language. MICoBot first selects an abstract verbal action from this space, then translates it into a natural language utterance to negotiate with the human—conveying its requests and the assistance it requires for successful collaboration. This involves reasoning over asymmetric human and robot physical capabilities to devise collaboration strategies that maximize task success while minimizing human effort.

Collaborative Task Definition and Problem Statement. We assume the collaborative task is defined by a task plan of length T, known to both agents and represented as a sequence of unassigned **physical** action primitives, $[a_{p,0}, ..., a_{p,T-1}]$, such as [(`pick-and-place(box, table)`, ..., `close(box)`], obtained from the task instructions or off-the-shelf task planner. To complete the manipulation task while minimizing human effort, the system must allocate steps of the plan between the two agents—negotiating with the human through robot-initiated dialog to suggest assignments, adapting to human preferences through human-initiated dialog, and ultimately executing its assigned physical actions. At each step t, the system must compute the best allocation of the remaining steps of the plan, $G = [g_t, ..., g_{T-1}]$, where $\forall t, g_t \in \{H, R\}$. The optimal allocation G^* maximizes the expected task success probability while minimizing total human effort. These objectives are inherently competing: a policy focused solely on maximizing success might allocate all steps to the human (assumed to be perfectly competent); conversely, minimizing human effort alone would assign all steps to the robot, even when it may be incapable of completing certain steps. The optimization also incorporates constraints conveyed through the mixed-initiative dialog history, such as task allocation requests or proposed task splits. The resulting allocation G^* determines whether the robot executes the current step (R) or negotiates with the human for assistance (H).

IV. MICoBot: Mixed-Initiative Collaborative Robot

A. Collaborative Task Allocation as Optimization.

A helpful physical collaborator must aim for task success with minimal human effort while adhering to human preferences expressed in dialog (i.e., for certain steps to be done by a certain agent). Therefore, we formulate our objective for collaborative task allocation as a constrained optimization problem, where constraints are updated based on dialog exchanges. To avoid a complex multi-objective

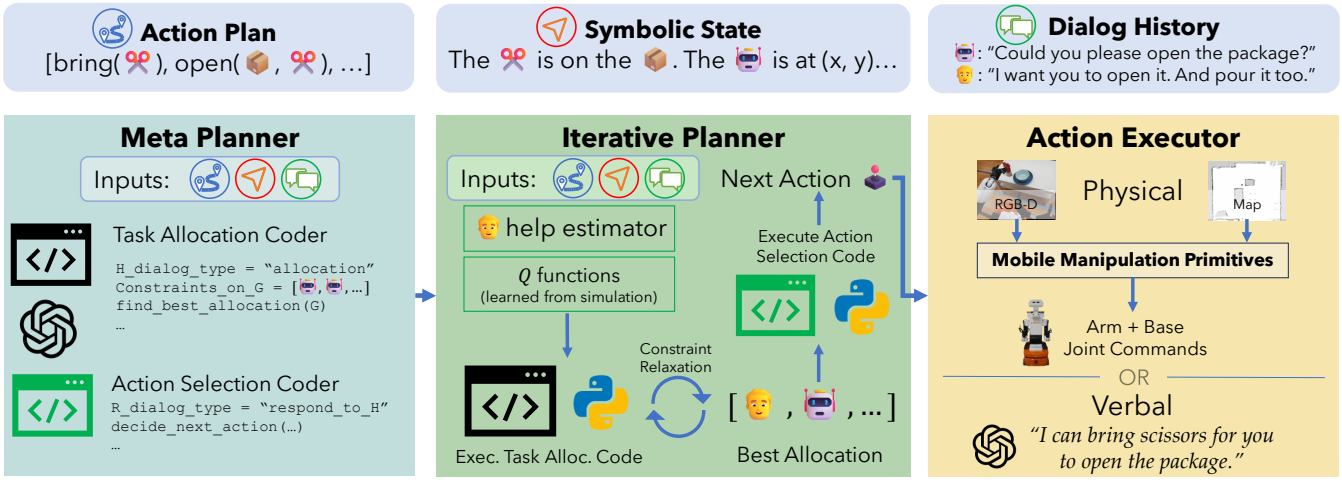

Fig. 4: MICoBot consists of 3 decision-making modules: a meta-planner that produces a collaborative strategy expressed through adaptive planning code, a planner that executes the code and optimizes our objective (Eq. 1) to decide the next primitive action, and an action executor that outputs the low-level pose trajectory or verbal utterance to say to the human.

formulation, we combine success probability and effort into a single Q-value by building on prior work on temporal distances in RL [42]. Then, to allocate task steps, MICoBot compares robot and human Q-values.

We assume each task step is executed by a multi-task policy π that performs continuous low-level control at a fixed control frequency. In this low-level MDP (distinct from the high-level task MDP described in Sec. III), we define the reward as $r = -1$ per time step until the skill completes or times out, at which point $r_{termination} = 0$. A well-trained Q-function, $Q : o_t \times a_t = (\omega_t, \theta_t) \mapsto \mathbb{R}$ with a discount factor of 1, thus represents the **negative expected number of timesteps** until skill completion from a given state. For a perfectly competent agent (i.e., human), this corresponds to the average timesteps required to perform the action. For an imperfect agent that may fail, the Q-function reflects a weighted expectation over both successful and failed outcomes—where failure contributes a significant timestep penalty (timeout) weighted by its probability. We assign each agent a distinct Q-function: Q_R for the robot and Q_H for the human. These agent-specific Q-functions thus provide a unified, interpretable cost metric for comparing step allocations, jointly capturing both execution time (effort) and likelihood of success.

However, directly optimizing step allocation using only these two Q-functions diverges from realistic human-robot collaboration scenarios in three ways: (1) human and robot effort are valued equally, ignoring the higher worth of human time and attention; (2) the human is assumed to always comply with robot-initiated requests, overlooking variability in willingness or availability; and (3) human-initiated requests or preferences are not considered, limiting the system's adaptability to human intent. To address (1), we introduce a *human-effort factor*, α, a ratio valuing human effort to robot effort. To address (2), human Q-values are adjusted with an inferred probability, $p_{H,t}$, of the human agreeing to perform action $a_{H,t} = \omega_t(\theta_t)$ when asked. For less cooperative users, this probability lowers the expected success of $a_{H,t}$, effectively increasing the magnitude of the negative Q-value due to potential human refusal. To address (3), we enforce constraints, $C_1, \ldots, C_n$, extracted from human-initiated dialog—such as explicit requests to perform specific steps themselves or delegate them to the robot. Altogether, we propose the following objective to find the optimal task allocation G^*:

$$\max_{g_t,\ldots,g_T} \sum_{t}^{T-1} \left(\mathbb{1}_{g_t=H} \cdot \frac{\alpha}{p_{H,t}} + \mathbb{1}_{g_t=R} \right) Q_{g_t}(s_t, a_t), \quad (1)$$
$$s.t. \quad C_1, \ldots, C_n \text{ are satisfied}$$

that minimizes expected time-to-success and human effort.

B. MICoBot Framework

MICoBot is a three-level framework (Fig. 4) that includes 1) a meta-planner that processes human dialog and generates a collaborative strategy expressed in code, 2) an iterative planner that updates planning state variables and allocates and decides the next action to perform by executing the code, and 3) an action executor that carries out the action primitive, either through low-level physical actions or by formulating a dialog utterance to communicate to the human.

L1: Meta-planner. The meta-planner produces adaptive planning code that dictates the overall strategy for L2 and L3 to follow. Based on the most recent human dialog, the current symbolic state of the world, the task plan, and approximately 15 in-context learning (ICL) examples, it generates two pieces of code: first, **task allocation** code to adapt the optimization computation, such as to map human dialog into additional constraints, and second, **action selection** code for how to choose the next action, such as whether to engage in additional dialog, proposals to split up the task with the human, or negotiation rounds before proceeding further with physical steps in the plan. The meta-planner is implemented

as an LLM-based (GPT-4o) coder, and prompts can be found on our project website.[1]

L2: Iterative Planner. The iterative planner runs code from the meta-planner (L1) to make two key decisions: whether to initiate dialog and which verbal or physical action to perform. L2 runs in two stages. *First*, it performs constrained optimization to find the best task allocation for the remaining task steps. To do this, we evaluate Eq. 1 under all possible task allocations fulfilling the constraints. Initially, the planner attempts to incorporate all constraints from the mixed-initiative dialog history. If no feasible allocation is found (e.g., human asked the robot perform a step it is incapable of), the planner iteratively relaxes the most recent constraint, and the robot verbally explains its incapability.

To evaluate Eq. 1, MICoBot requires accurate Q-functions that capture each agent's expected effort and success probability on each task step. To collect data to learn the robot's Q-function (Q_R), we use the OmniGibson simulator [43], configured with a coarse model of the real-world task, environment, and action primitives, recording both completion times and success rate. We train a supervised network as Q_R that predicts the expected timesteps for an action primitive a to succeed from a given symbolic state o. Conditioning Q_R on symbolic states minimizes the sim-to-real gap when we deploy the function to our real-world setting. When estimating the human's Q-function (Q_H), we assume perfect competence (i.e., no execution failures). Thus, we simply obtain time estimates for each step from an LLM predicting how long a human needs to execute action $a_t = \omega_t(\theta_t)$, plus a travel time estimate based on human-object distances.

To adapt to changing human helpfulness, MICoBot estimates the probability of human assistance at the current t-th timestep, $p_{H,t}$, using an LLM-based sentiment analysis over prior human-robot dialog. This enables MICoBot to adapt to temporally-changing user sentiments. After deciding the optimal task allocation, the second stage of L2 executes meta-planner action selection code to generate the optimal action $a = \omega(\theta)$ to execute: a physical mobile manipulation primitive ω to perform a task step on objects specified in θ, or a verbal primitive ω to initiate dialog to ask for help, propose splitting up steps, or respond to human-initiated dialog regarding specific task steps specified in θ.

L3: Action Executor. The action executor performs the action primitive selected by the planner (L2). For *physical actions*, it generates a trajectory for navigation and arm movement to reach the location of and manipulate the target object while avoiding obstacles. We build a pipeline similar to [44] that uses the `move_base` ROS package for navigation path planning over a 2D occupancy map, and Grounding DINO [45] to segment the target object from an open-world scene based on the object specified in θ. We backproject segmented image pixels from RGB-D camera data into a 3D point cloud to identify graspable or placeable points in the robot's workspace that the arm reaches through inverse kinematics (IK). For *verbal actions*, an LLM generates free-form natural language utterances to communicate with the human based on the dialog intent ω and verbal action parameters θ decided in L2. The LLM uses 10 in-context-learning examples to generate language grounded in the context of the task and dialog.

Hierarchical Plan. To streamline communication for long-horizon task plans, MICoBot groups adjacent low-level steps into semantically meaningful abstract actions that can be discussed more succinctly with the human. The system only descends to a finer-grained level of detail during negotiation over low-level step assignments, which reduces the frequency and complexity of dialog, resulting in more efficient and user-friendly communication.

V. Experimental Evaluation

We evaluate MICoBot in the real-world, on a Tiago mobile manipulator working with 18 unique human participants on household tasks, and in simulation, on a collaborative framework we developed atop Mini-Behavior gridworld [4]. In our simulation framework, a robotic agent collaborates with a simulated human with parametrizable helpfulness and mood-varying dialog, which allows for larger-scale experimentation and controlled comparisons across methods across a wider range of human behavior and dialog dynamics. A successful robotic collaborator must achieve task success (our primary evaluation metric) while minimizing human effort (our secondary metric). We also report **subjective measures of robot behavior**, including user satisfaction, preference rankings, and Likert-scale ratings.

Environment. In the real-world, we perform our experiments in a mock apartment with a kitchen and living room area with commonplace furniture. In all of our tasks, the robot and human work together on opposite sides of a coffee table. Simulating a household setting, the participant spends nearly all of their time on the couch, where they can do their personal (i.e., non-task-related) work. The human can be as inactive or proactive as they wish in performing physical and verbal actions as defined in Section III (though we continue the trial if they initiate dialog beyond the scope). Each human user study consists of two 20-30 minute trials, in which they collaborate with both our method and a pure LLM baseline, ordered randomly. All trials **terminate** under any of the following conditions: a primitive fails irrecoverably, $4T$ steps have elapsed for a plan of length T, an infeasible step is allocated to the robot twice consecutively, or the human refuses twice to perform a step infeasible for the robot.

Skills. To perform long-horizon household tasks, the robot has access to several mobile-manipulation action primitives. `pick_place_mobile(obj, place_loc)` moves to `obj` and places it atop `place_loc`, another object in a potentially different room. `pour(obj, cont)` travels to `obj` and pours its contents into `cont`. Finally, `fold(obj)` folds down box flaps.

To initiate and respond within mixed-initiative dialog, the robot uses the following open-vocabulary verbal action primitives for dynamic collaboration with the human: `ask_for_human_help` on a `step` the human is best suited to perform, `propose_split` to split `steps` with

	Pour Package in Bowl $n=6$		Assemble Toy Car $n=6$		Pack Gift Box $n=6$		Average $n=18$	
	MICoBot	LLM	MICoBot	LLM	MICoBot	LLM	MICoBot (ours)	LLM
Entire Task Success Rate (%, ↑)	**100**	83	**67**	0	**67**	0	$\mathbf{77.8 \pm 15.7}$	27.8 ± 39.3
% of task steps completed (↑)	**100**	93	**94**	31	**88**	50	$\mathbf{93.8 \pm 5.1}$	58.2 ± 26.0
% of steps performed by Human	27	29	60	5	35	21	40.5 ± 14.2	18.2 ± 9.7
% Users Preferring ... (↑)	**67**	33	**100**	0	**67**	33	**77.8**	22.2
Communicative ability (↑, /5)	3.7	**3.8**	**4.3**	1.3	**2.8**	2.3	$\mathbf{3.6 \pm 1.0}$	2.5 ± 1.4
Awareness of its Limitations (↑, /5)	**3.3**	**3.3**	**3.7**	1.2	**4.2**	2.5	$\mathbf{3.7 \pm 1.4}$	2.3 ± 1.6
Overall Satisfaction working w/ Robot (↑, /5)	**3.8**	3.7	**3.5**	1.5	**3.5**	2.5	$\mathbf{3.6 \pm 0.8}$	2.6 ± 1.4

TABLE I: Comparison between MICoBot (ours) and the LLM baseline across three real-world tasks on both objective (top 3 rows) and subjective (bottom 4 rows) metrics. Ratings out of 5 are on the Likert scale. Through more effective task allocation and communication, MICoBot achieves much higher task success rates and overall user satisfaction.

the human, `explain_incapability` to ask the human to perform a `step` that the robot can't perform, and `respond_to_human` to `accept/reject` requests the robot is capable/incapable of executing.

Baselines. Because multiple components of our method are powered by LLMs, we compare our approach to a pure LLM baseline (**LLM**) given the same information as our meta-planner: symbolic state, dialog history, task plan, and α human-robot effort tradeoff factor. The LLM baseline is also provided with a list of the robot's available skills and assumes that the human always successfully completes a step once they agree to perform it. The LLM baseline is prompted to produce a plan allocation G that primarily optimizes for task success and secondarily minimizes human effort.

To control for the amount of human effort in the user studies, we compute an additional random allocation baseline that does not involve a human participant, **RECB** (random effort-controlled baseline). Let p_c denote the proportion of steps done by the human in our method's trials. RECB randomly allocates the current step to the human with probability p_c and assumes a perfectly helpful human and oracle robot primitives with 100% success rate.

In simulation, we additionally compare against an **RL** baseline (hierarchical task allocator + robot policy) and a naive **Random** baseline allocating either agent (with probability 50%) to perform the next step.

Ablations. To measure the importance of mixed-initiative dialog, we perform the following ablations in simulation: **H-init** and **R-init**, where the human or the robot alone, respectively, can initiate any dialog. We further ablate components of MICoBot in simulation by running it **w/o P_H** (no $p_{H,t}$ estimation) and **w/o Plan Hierarchy** (where our method talks to the human in terms of granular, low-level steps instead of more understandable, high-level subtasks).

Tasks. We perform user studies on 3 real-world tasks, each with 6 users for a total of 18 unique participants. **(1) Pour package into bowl**: bring the bowl, package, and scissors from the kitchen, cut open the package, and pour it into the bowl. **(2) Assemble toy car**: bring the car parts, wheels, and drill from the shelf to the coffee table, drill in the wheels, switch the drill bit, and finally drill in the windows and seats. **(3) Pack gift box**: fold the gift box, put tissue wrapping paper and a toy car in the box, close the lid, wrap ribbons, and tape down a gift bow. Each task is 5 to 8 mobile manipulation steps long and requires varying degrees of human involvement.

Experimental analysis. Our experiments are designed to answer the following research questions:

(1) Does our method achieve the best trade-off between task success and minimizing human effort? In our real-world user study (Table I), MICoBot achieves a 78% task success rate compared to 28% for the LLM baseline (statistically significant with p-value 0.007 under Fisher's exact test). Additionally, MICoBot achieves a 94% task step completion rate compared to the baseline's 58% (statistically significant with p-value 0.002 under the Wilcoxon-signed-rank test). MICoBot understood its own limitations (through affordance functions trained in simulation), and was hence better at leveraging human assistance effectively on the steps it was ill-suited to perform. The LLM baseline tended to prioritize minimizing human effort over task completion by allocating the robot multiple steps it was incapable of, since the LLM lacked an understanding of the robot's affordances. Our method elicited more human effort than the baseline (40% vs 18%), so to control for the amount of human effort received, we compare our method to RECB in Figure 5. Despite RECB assuming oracle robot primitives and a perfectly cooperative human, our method still significantly outperforms it by more effectively balancing between success and human workload.

(2) How do users feel about working with our system? The A/B blind preference test in Table I shows that 78% of users preferred our method over the LLM baseline. Our method also significantly outperformed the baseline in user scores on overall satisfaction, communicative ability, and capability in asking for a suitable amount of help (statistically significant under the Wilcoxon-signed-rank test with p-values ranging from 0.007 to 0.024; see Figure 6). In contrast, the LLM baseline often failed to express when it needed help and was unwilling to reject human requests it could not fulfill, leading to over-promises and task failures. A representative dialog exchange, available on our project site,[1] shows MICoBot successfully persuading an initially reluctant user to perform a step the robot was incapable of executing.

(3) Is mixed-initiative dialog critical to our method's

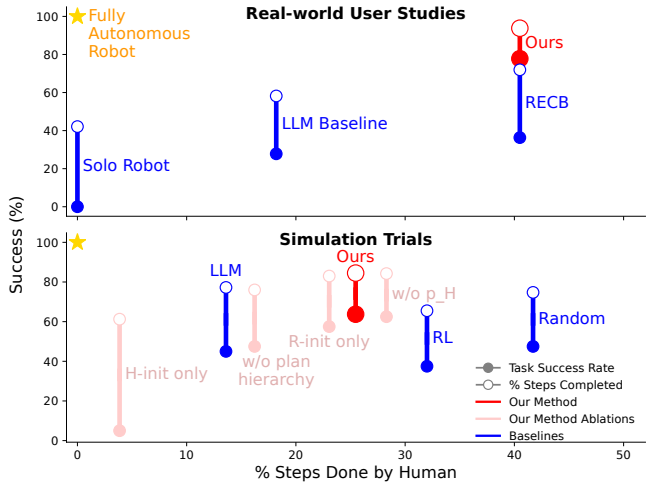

Fig. 5: In both **real-world** user studies **(top)** and **simulation trials** with a simulated human **(bottom)**, our method (red) demonstrates the best tradeoff in achieving task success (y-axis) for a given amount of human effort (x-axis) than baselines (blue) and our method's ablations (pink).

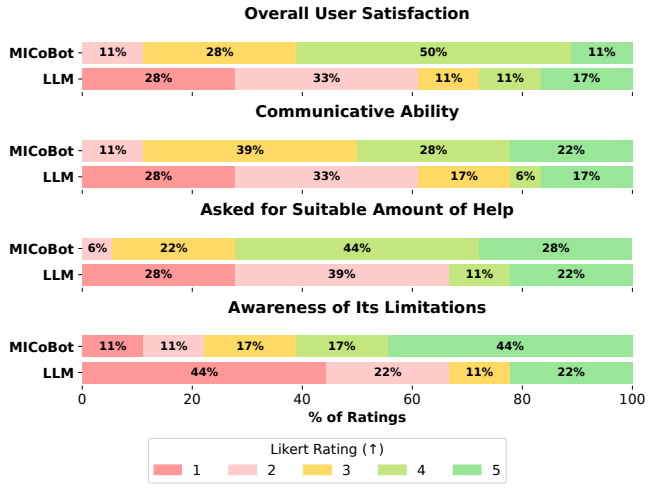

Fig. 6: Our method substantially outperforms the pure LLM baseline in user ratings averaged over all $n = 18$ participants.

performance? Figure 5 (bottom) shows that our full method outperforms both ablated variants that restrict dialog to single-initiative modes: robot-only initiation (R-init) and human-only initiation (H-init). H-init performs especially poorly, as it prevents the robot from requesting help for steps it cannot execute. R-init performs slightly worse than the full method because it does not allow the human to proactively initiate dialog and assist when appropriate. These results underscore the importance of mixed-initiative dialog in enabling flexible, robust human-robot collaboration.

In real-world user studies, MICoBot engaged in 2.4 dialog initiative shifts per trial, compared to the LLM baseline's 1.1. This enabled MICoBot to boost human acceptance of help requests from 55% to 86%. The LLM baseline made far fewer help requests per trial (0.9 vs. MICoBot's 2.9) and achieved a smaller acceptance increase (70% to 75%). This demonstrates mixed-initiative dialog is critical to collaborative discussion and task success.

VI. Conclusion

We proposed MICoBot, a real-world robotic system that improves collaboration on long-horizon mobile manipulation tasks through mixed-initiative dialog with humans. Our work unifies two previously unconnected lines of research: mixed-initiative dialog and HRI. To this end, we formulated a novel optimization function and robotic framework using mixed-initiative dialog as a rich interface for task allocation to maximize task success while minimizing human effort and complying with verbally-expressed human preferences. Real-world user studies with 18 human participants and extensive trials in simulation demonstrate the efficacy, adaptability, and user satisfaction of our method across a diverse range of human physical and verbal behavior.

VII. Limitations and Future Work

MICoBot represents our pioneering effort on facilitating mixed-initiative human-robot interaction through mixed-initiative natural-language dialog. While we focused on delegating steps for long-horizon manipulation tasks in a manner that maximizes task success and minimizes human effort, we believe this paper opens up exciting new avenues for future work. These include enabling both agents to learn to provide and incorporate spatial-temporal feedback to each other while performing a task, share relevant task information in an imperfect-information setting, and replan and redefine a task as necessary, all through mixed-initiative dialog.

MICoBot has a number of limitations. First, it assumes that the human and robot work sequentially, and cannot handle cases where a robot and human wish to collaborate simultaneously on the same step in the plan, such as if the robot holds a roll of tape and the human cuts from it. Second, MICoBot assumes a fixed plan with a predetermined ordering of steps, where the human has a general, high-level understanding of the plan (but not the low-level steps that the robot plans over). Our method could be improved with a more nuanced definition of "effort" beyond our time-based metric. Finally, $p_{H,t}$ prediction can be improved, such as by processing tone-of-voice and facial expressions, to enable producing more emotionally understanding dialog, which can improve task success and user satisfaction.

Acknowledgments

We thank Rutav Shah and Arpit Bahety for robot support, and Ben Abbatematteo for manuscript edits. Thanks also to our study participants and RobIn and ML lab members for feedback. This work was funded by DARPA TIAMAT HR0011-24-9-0428, Amazon Award, Emerson Electric, and the UT Austin Graduate School Fellowship.

References

[1] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," *NeurIPS*, 2022.

[2] J. Achiam *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[3] M. Selvaggio, M. Cognetti, S. Nikolaidis, S. Ivaldi, and B. Siciliano, "Autonomy in physical human-robot interaction: A brief survey," *IEEE RA-L*, 2021.

[4] E. Jin *et al.*, "Mini-behavior: A procedurally generated benchmark for long-horizon decision-making in embodied ai," *arXiv preprint 2310.01824*, 2023.

[5] J. R. Carbonell, "Ai in cai: An artificial-intelligence approach to computer-assisted instruction," *IEEE Transactions on Man-Machine Systems*, vol. 11, no. 4, pp. 190–202, 1970.

[6] J. Allen, C. Guinn, and E. Horvtz, "Mixed-initiative interaction," *IEEE Intelligent Systems and their Applications*, vol. 14, no. 5, pp. 14–23, 1999.

[7] J. Chu-Carroll, "MIMIC: An adaptive mixed initiative spoken dialogue system for information queries," ACL, 2000.

[8] S. Wu *et al.*, "Collabllm: From passive responders to active collaborators," in *ICML*, 2025.

[9] Y. Deng, W. Lei, W. Lam, and T.-S. Chua, "A survey on proactive dialogue systems: Problems, methods, and prospects," *arXiv preprint arXiv:2305.02750*, 2023.

[10] X. Yu, M. Chen, and Z. Yu, *Prompt-based monte-carlo tree search for goal-oriented dialogue policy planning*, 2023. arXiv: 2305.13660 [cs.CL].

[11] M. Chen, X. Yu, W. Shi, U. Awasthi, and Z. Yu, "Controllable mixed-initiative dialogue generation through prompting," *arXiv preprint arXiv:2305.04147*, 2023.

[12] Y. Deng, W. Zhang, W. Lam, S.-K. Ng, and T.-S. Chua, *Plug-and-play policy planner for large language model powered dialogue agents*, 2024. arXiv: 2311.00262 [cs.CL].

[13] K. Qian *et al.*, "Database search results disambiguation for task-oriented dialog systems," *NAACL*, pp. 1158–1173, 2022.

[14] Y. Deng, L. Liao, L. Chen, H. Wang, W. Lei, and T.-S. Chua, *Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration*, 2023. arXiv: 2305.13626 [cs.CL].

[15] M. Chen, R. Sun, S. O. Arık, and T. Pfister, *Learning to clarify: Multi-turn conversations with action-based contrastive self-training*, 2024. arXiv: 2406.00222 [cs.CL].

[16] J. Thomason *et al.*, "Improving grounded natural language understanding through human-robot dialog," *ICRA*, pp. 6934–6941, 2019.

[17] H. Wang *et al.*, *Mosaic: A modular system for assistive and interactive cooking*, 2024. CoRL: 2402.18796 (cs.RO).

[18] Z. Mandi, S. Jain, and S. Song, "Roco: Dialectic multi-robot collaboration with large language models," *ICRA*, 2024.

[19] X. Feng *et al.*, "Large language model-based human-agent collaboration for complex task solving," *EMNLP Findings*, pp. 1336–1357, 2024.

[20] A. Suhr *et al.*, *Executing instructions in situated collaborative interactions*, 2022. arXiv: 1910.03655 [cs.CL].

[21] N. Kojima, A. Suhr, and Y. Artzi, *Continual learning for grounded instruction generation by observing human following behavior*, 2021. arXiv: 2108.04812 [cs.CL].

[22] D. I. A. Team *et al.*, *Creating multimodal interactive agents with imitation and self-supervised learning*, 2022. arXiv: 2112.03763 [cs.LG].

[23] Q. Gao *et al.*, "Alexa arena: A user-centric interactive platform for embodied ai," *NeurIPS*, vol. 36, 2023.

[24] A. Z. Ren *et al.*, "Robots that ask for help: Uncertainty alignment for large language model planners," *CoRL*, 2023.

[25] A. Bennetot, V. Charisi, and N. Díaz-Rodríguez, *Should artificial agents ask for help in human-robot collaborative problem-solving?* 2020. arXiv: 2006.00882 [cs.LG].

[26] R. A. Knepper, T. Layton, J. Romanishin, and D. Rus, "Ikeabot: An autonomous multi-robot coordinated furniture assembly system," in *2013 IEEE International Conference on Robotics and Automation*, 2013, pp. 855–862.

[27] M. Veloso, J. Biswas, B. Coltin, and S. Rosenthal, "Cobots: Robust symbiotic autonomous mobile service robots," in *IJCAI, 2015*, Buenos Aires, Argentina: AAAI Press, 2015, pp. 4423–4429.

[28] D. L. Chen, J. Kim, and R. J. Mooney, "Training a multilingual sportscaster: Using perceptual context to learn language," *Journal of Artificial Intelligence Research*, vol. 37, pp. 397–435, 2010.

[29] B. Mutlu, J. Forlizzi, and J. Hodgins, "A storytelling robot: Modeling and evaluation of human-like gaze behavior," in *2006 6th IEEE-RAS International Conference on Humanoid Robots*, 2006, pp. 518–523.

[30] S. Cascianelli, G. Costante, T. A. Ciarfuglia, P. Valigi, and M. L. Fravolini, "Full-gru natural language video description for service robotics applications," *IEEE RA-L*, vol. 3, no. 2, pp. 841–848, 2018.

[31] D. A. Few, D. J. Bruemmer, and M. C. Walton, "Improved human-robot teaming through facilitated initiative," in *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, 2006, pp. 171–176.

[32] M. Natarajan, C. Xue, S. van Waveren, K. Feigh, and M. Gombolay, "Mixed-initiative human-robot teaming under suboptimality with online bayesian adaptation," *AAMAS*, pp. 1398–1407, 2024.

[33] J. Bishop *et al.*, "Chaopt: A testbed for evaluating human-autonomy team collaboration using the video game overcooked!2," in *2020 Systems and Information Engineering Design Symposium (SIEDS)*, 2020, pp. 1–6.

[34] A. Rosero, F. Dinh, E. J. de Visser, T. Shaw, and E. Phillips, "Two many cooks: Understanding dynamic human-agent team communication and perception using overcooked 2," *AAAI-FSS*, 2021.

[35] R. Paleja, M. Munje, K. Chang, R. Jensen, and M. Gombolay, *Designs for enabling collaboration in human-machine teaming via interactive and explainable systems*, 2024. arXiv: 2406.05003 [cs.RO].

[36] S. Jiang and R. C. Arkin, "Mixed-initiative human-robot interaction: Definition, taxonomy, and survey," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2015, pp. 954–961.

[37] J. Baraglia, M. Cakmak, Y. Nagai, R. Rao, and M. Asada, "Initiative in robot assistance during collaborative task execution," in *HRI*, 2016, pp. 67–74.

[38] S. Vats, O. Kroemer, and M. Likhachev, *Synergistic scheduling of learning and allocation of tasks in human-robot teams*, 2022. arXiv: 2203.07478 [cs.RO].

[39] T. Yu, J. Huang, and Q. Chang, "Optimizing task scheduling in human-robot collaboration with deep multi-agent reinforcement learning," *Journal of Manufacturing Systems*, vol. 60, pp. 487–499, 2021.

[40] M. Faccio, I. Granata, and R. Minto, "Task allocation model for human-robot collaboration with variable cobot speed," *Journal of Intelligent Manufacturing*, 2024.

[41] S. Singh, A. Srikanthan, V. Mallampati, and H. Ravichandar, "Concurrent constrained optimization of unknown rewards for multi-robot task allocation," *RSS*, 2023.

[42] V. Myers, C. Zheng, A. Dragan, S. Levine, and B. Eysenbach, "Learning temporal distances: Contrastive successor features can provide a metric structure for decision-making," *ICML*, 2024.

[43] C. Li *et al.*, "BEHAVIOR-1k: A benchmark for embodied AI with 1,000 everyday activities and realistic simulation," in *CoRL*, 2022.

[44] R. Shah, A. Yu, Y. Zhu, Y. Zhu, and R. Martín-Martín, "Bumble: Unifying reasoning and acting with vision-language models for building-wide mobile manipulation," *ICRA*, 2025.

[45] S. Liu *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *ECCV*, 2024.