Slow the Dialogue, Not Just the Robot: Positive Friction for Reliable Grounding and Safe, Embodied Vision-Language Action

Anonymous ACL submission

Abstract

Embodied conversational robots must translate underspecified natural language commands into physical actions where mistakes can be costly or irreversible. Current LLM-based robot systems often act immediately: guessing missing referents; spatial relations; or motion constraints, leading to task failures and safety risks. As a response to this, we present PONDER, a dialogue architecture that operationalizes positive friction for embodied interaction: when the current visual context admits multiple plausible interpretations, the system inserts targeted clarification questions, explicit assumption statements, or brief confirmation pauses before execution. PONDER runs on a Misty II mobile robot, integrating speech input, a vision-language model, and conversational memory with navigation and perception actions. We carry out a user study, where positive friction increases task success from 18.8% to 89.6% and improves user ratings from 1.29 to 3.85 (5-point scale), at the cost of only an average of 1.14 additional dialogue turn. In addition, we verify our results in a simulated setup across diverse ambiguity types, where PONDER achieves 74.8% success versus 60.3% without friction and substantially outperforms zero-shot baselines (37.8 - 44.8%). We release an open-source Misty II implementation and our synthetic dialogue dataset to support reproducible research on embodied dialogue.

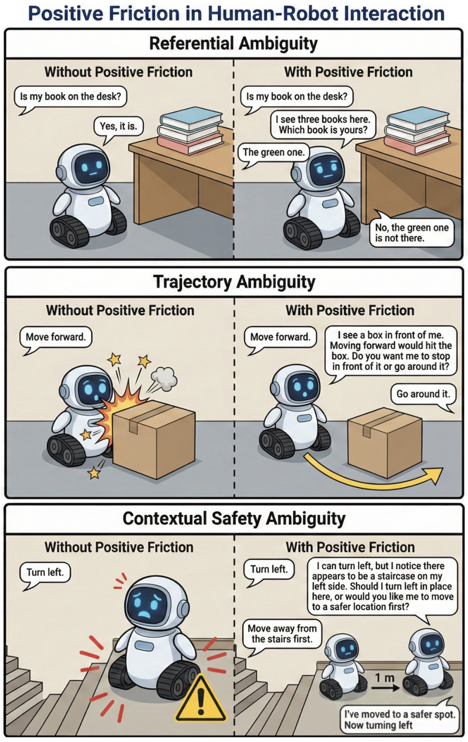


Figure 1: Examples of how positive friction mechanisms improve task outcomes across three ambiguity categories. Without friction, the robot executes commands immediately, leading to incorrect responses, collisions, or safety violations. With friction, the robot initiates targeted clarification dialogues to resolve ambiguities before physical action, demonstrating improved situational awareness and user collaboration.

1 Introduction

Recent advances in Large Language Models have enabled natural language interfaces for manipulating embodied agents, opening new possibilities for human-robot collaboration through dialogue (Ahn et al., 2022; Huang et al., 2023b). Yet grounding language to physical action remains fragile where human instructions are routinely underspecified, and robots operate under partial observability with uncertain perception. When a system commits to an interpretation too early, errors propagate into physical motion and can cause task failure, collisions, or unsafe behavior.

Ambiguity is especially pervasive in spatial commands. Utterances such as *"go to the plant," "put that on the table,"* or *"move forward a bit"* may be consistent with multiple referents, spatial relations, or feasible trajectories given the current scene. Many robot dialogue systems either guess missing details and act immediately, or ask generic follow-up questions that feel unnecessary and degrade interaction quality. This creates a tension between apparent efficiency (fewer turns) and reliability (correct, safe execution).

Dialogue theory offers a path through this tension. Successful communication depends on building common ground: participants make assumptions explicit and confirm mutual understanding before committing to consequential actions (Clark, 1996a; Clark and Brennan, 1991a). Recent work formalizes this idea as *positive friction*—deliberate dialogue moves that add small, targeted interaction cost to surface assumptions, probe ambiguities, and prevent downstream failures (İnan et al., 2025). Positive friction has shown gains in text-based task settings, but its effectiveness under real-time multimodal perception and physical execution has not been established. Embodied robots introduce constraints that make premature execution particularly costly: irreversible motion, safety-critical decisions, and visual grounding uncertainty.

We present PONDER, a positive-friction-oriented dialogue architecture for embodied conversational robots. PONDER treats ambiguity resolution as a first-class decision in the language-to-action loop. Given a spoken command and the robot's current visual observation, a vision-language model assesses whether multiple plausible interpretations exist. When ambiguity is detected, the system generates targeted clarification questions, explicitly states assumptions for confirmation, or pauses to request verification before executing navigation or perception actions. We implement PONDER on the Misty II mobile robot with speech recognition, a vision-language model, and conversational memory, enabling multi-turn dialogue that prioritizes common ground over turn minimization.

We evaluate PONDER in both physical and simulated settings. In a user study with 8 participants on tasks containing referential and trajectory ambiguities, positive friction increases task success from 18.8% to 89.6% and substantially improves user satisfaction, while requiring roughly one additional dialogue turn on average. To test robustness across a broader set of scenarios, we run 500-episode synthetic evaluations per condition using a structured world model and a vague-to-specific user protocol. PONDER achieves 74.8% task success compared to 60.3% without friction and outperforms zero-shot baselines by a large margin. Together, these results show that deliberate conversational slowdown can improve both reliability and user experience in embodied interaction.

Our contributions are threefold:

1. We identify and operationalize ambiguity types central to embodied dialogue: referential; trajectory; contextual safety; and implicit preconditions.
2. We introduce PONDER, an embodied positive-friction architecture deployed on a physical conversational robot, integrating voice, vision-grounded reasoning, and multi-turn memory to decide when to clarify versus act.
3. We release an open-source implementation for Misty II and a synthetic dialogue dataset to support reproducible research on friction-based embodied dialogue policies.

2 Related Work

Vision-Language-Action Models LLMs have enabled significant advances in robotic control through natural language (Zitkovich et al., 2023; Kim et al., 2025; Black et al., 2024). Foundational approaches ground language in affordances (Ahn et al., 2022), generate executable code (Liang et al., 2023), or incorporate environmental feedback (Huang et al., 2023b). More recent work explores LLMs as high-level planners (Song et al., 2023; Lin et al., 2023) or for direct motor control (Collaboration et al., 2025). The ReAct framework (Yao et al., 2023) introduced interleaved reasoning and acting, extended by ReSpAct (Dongre et al., 2025) to embodied settings with explicit communication. Despite advances in multimodal architectures (Shridhar et al., 2022; Jiang et al., 2023; Chi et al., 2023) and grounded decoding (Huang et al., 2023a), grounding spatial language to physical referents remains challenging (Tellex et al., 2020; Bisk et al., 2020). Benchmarks like ALFRED (Shridhar et al., 2020) and TEACh (Padmakumar et al., 2022) highlight persistent difficulties in resolving ambiguous references. Most systems optimize for task completion and rarely address the ambiguity resolution process itself; our work introduces a dialogue layer that strategically slows interactions to build common ground before action execution.

Common Ground and Ambiguity in Dialogue Our theoretical foundations draw from Clark's work on common ground (Clark, 1996b; Clark and Brennan, 1991b), which established that successful communication requires collaborative effort through presentation and acceptance phases. Traum (Traum, 1994) operationalized grounding

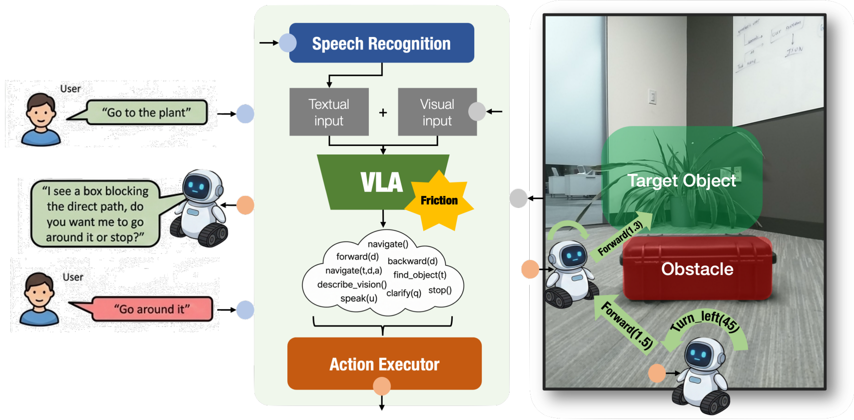


Figure 2: Overview of our proposed approach. Given a spoken natural language command and a visual observation from the robot's camera, both inputs are passed to a vision-language model (VLM) augmented with our positive friction mechanism. The VLM first analyzes whether the command contains any ambiguity given the current visual context. When ambiguity is detected, the system initiates targeted clarification dialogue rather than guessing; when the command is unambiguous, it proceeds directly to action execution.

for computational systems, and Chai and colleagues (Chai et al., 2014, 2017) extended this framework to human-robot interaction. Ambiguity pervades natural language at multiple levels (Poesio and Artstein, 2005; Piantadosi et al., 2012); benchmarks including AmbigNQ (Min et al., 2020) and AmbiQT (Bhaskar et al., 2023) provide taxonomies for question answering and code generation. Recent work shows LLMs can generate clarification questions but struggle to identify when clarification is truly needed (Deng et al., 2023; Zhang et al., 2024b).

Positive Friction and Trustworthy Human-Robot Interaction Positive friction (İnan et al., 2025; Obiso et al., 2025) formalizes intentional dialogue moves that slow interaction to reveal assumptions and probe ambiguities. Related work connects common ground maintenance to conversational success (Sarkar et al., 2025) and explores trust calibration in AI-assisted decisions (Srinivasan and Thomason, 2025). Safe human-robot interaction requires appropriate uncertainty handling (Lasota et al., 2017), addressed through active preference learning (Sadigh et al., 2017; Bıyık et al., 2024) and interactive imitation (Ross et al., 2011). Work on collaborative reference (Clark and Wilkes-Gibbs, 1986; Haber et al., 2019) demonstrates that grounding processes ultimately reduce total collaborative effort. Our contribution integrates dialogue theory with physical manipulation, showing that positive friction mechanisms serve both collaborative understanding and safety in embodied interaction.

While prior work has advanced vision-language-action models, dialogue grounding, and positive friction independently, no existing system integrates these components for physical robot interaction. Text-based positive friction studies operate in simulated environments where actions are reversible and perception is symbolic. Embodiment changes the problem qualitatively: commands must be grounded in noisy visual observations, execution is irreversible, and failures carry physical consequences. Our work bridges this gap by deploying positive friction on a physical robot under real-time constraints.

3 Our Approach

This section describes our system architecture, action space, task definitions, and evaluation methodology. Figure 2 provides an overview.

Problem Definition We consider a dyadic human-robot interaction where an instructor H gives natural language commands to guide a robot R toward a goal state g. At timestep t, the instructor produces utterance u_t while the robot perceives observation o_t. A command is *ambiguous* when u_t admits multiple valid interpretations given o_t.

Formally, given ambiguous command u_t and observation o_t, let $\mathcal{I}(u_t, o_t) = \{i_1, i_2, \ldots, i_k\}$ denote plausible interpretations, where $|\mathcal{I}| > 1$ indicates ambiguity. The **disambiguation task** is to identify the intended interpretation $i^* \in \mathcal{I}$ before action execution. A naive approach selects arbitrary $i \in \mathcal{I}$ and executes immediately, risking failure when $i \neq i^*$. We instead propose dialogue-based disambiguation: generating clarification query q_t to reduce $|\mathcal{I}|$. Given response r_t, the robot refines its interpretation set: $\mathcal{I}' = \textsc{Refine}(\mathcal{I}, q_t, r_t)$. This loop iterates until $|\mathcal{I}'| = 1$, then executes $a_t = \pi(i^*, o_t)$. We operationalize this through *positive friction*, which provides principled strategies for effective clarification.

PONDER Architecture We introduce *PONDER: Positive-Friction Oriented Natural Dialogue for Embodied Reasoning*. Our system (Equation 1) takes initial configuration $\mathcal{W}_0$ and goal $\mathcal{G}$, returning terminal state $\mathcal{W}^*$. At turn t, the robot obtains partial observation via $\text{POV}(\mathcal{W}_t)$ and invokes a VLM: $r_t = \text{VLM}(u_t, \text{POV}(\mathcal{W}_t), H_t)$, where H_t is conversation history. The recursion branches on response type: if the goal is satisfied, return current state; if $r_t = \texttt{EXEC}$, execute and update $\mathcal{W}_t'$; if $r_t = \texttt{CLAR}$, generate clarification, elicit response u_{t+1}, and recurse until unambiguous execution is possible.

$$\mathcal{W}^* = \textsc{ponder}(\mathcal{W}_0, \mathcal{G}) \tag{1}$$

$$= \begin{cases} \mathcal{W}_t & \text{if } \textsc{goal}(\mathcal{W}_t, \mathcal{G}) \\ \textsc{ponder}(\mathcal{W}_t, \mathcal{G}) & \text{if } r_t = \textsc{exec} \\ \textsc{ponder}(\mathcal{W}_t \cup \{u_{t+1}\}, \mathcal{G}) & \text{if } r_t = \textsc{clar} \end{cases} \tag{2}$$

PONDER uses a frozen, pre-trained VLM (GPT-5 nano) within a hand-designed dialogue controller. The clarify-versus-execute decision emerges from in-context prompting, and friction behavior can be toggled via prompt modification without retraining.

3.1 Tasks, the World Model & Actions

Tasks We design scenarios focusing on four ambiguity types common in embodied spatial dialogue (Min et al., 2020; Zhang et al., 2024b; Inan et al., 2025). Each scenario specifies robot pose, object coordinates and properties, environmental hazards, and task goals with expected ambiguity types. Table 1 summarizes the distribution and provides examples.

- *Referential:* underspecified object references when multiple candidates exist (Min et al., 2020).
- *Trajectory:* ambiguous path planning around obstacles (Anderson et al., 2018; Chen et al., 2024).
- *Contextual Safety:* environmental dangers requiring caution (Hu et al., 2024).
- *Implicit Precondition:* target objects outside the robot's current field of view with no prior context (Zhang et al., 2024a).
- *Multi-category:* multiple ambiguity types in a single task.

Ambiguity Type	**Example Task**	**#**
Referential	Go to the plant with multiple instances of box	10
Trajectory	Go to the table, chair blocking path to table	10
Implicit Precond.	Go back to the room(no room visible)	8
Contextual Safety	Move forward on the edge of a table	8
Multi-category	Navigate and find object	10

Table 1: Distribution of our defined task scenarios by ambiguity type. Scenarios test whether friction-enabled dialogue improves task success across diverse ambiguous tasks.

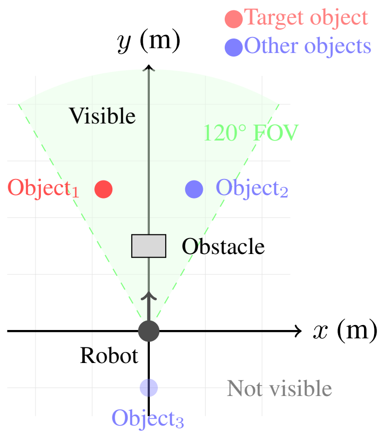


Figure 3: A visual example of our proposed world model. Robot at origin with 120° FOV (shaded). Red circles indicate target objects; blue circles indicate other objects. Objects within the FOV are visible; Object$_3$ (faded) behind the robot is not visible. Gray rectangle represents an obstacle for collision detection.

World Model To enable precise evaluation and reproducible experiments, we implement a 2-dimensional coordinate-based world model that tracks robot state and object positions (Figure 3). The world model maintains:

- **Robot State**: Position (x, y) in meters and orientation θ in degrees
- **Object Registry**: Objects with coordinates, properties, and dynamic visibility status
- **Hazard Zones**: Edge hazards and obstacles with collision detection
- **Action History**: Sequential record of all executed actions

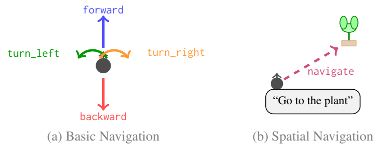
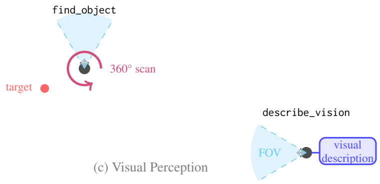
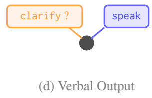
Category	Action Type	Example	Action Diagram
Navigation	forward(d) backward(d) turn_left(deg) turn_right(deg) navigate(t,d,a)	User: "Go to my office" Robot: clarify("I see multiple office rooms in view, which one is yours?.") User: 'The one with the light on." Robot: navigate(office, 3.0, -45) Robot: speak("I've reached your office") User: "Okay, now help me find my bag" Robot: find_object(bag)	
Perception	find_object(t) describe_vision()	Robot: clarify("I found multiple bags, could you describe the appearance of your bag?") User: "It's a small white bag" Robot: speak("I found your small white bag! It's on the purple chair to my left.") User: "Thank you, now please go to the table."	
Utterance	speak(u) clarify(q)	Robot: clarify(I see a box in front of me blocking the direct path to the table, would you want me to go around it or stop?) User: "Go around it."	
Control	stop()	Robot: speak("I'll navigate around the left side where there's more clearance.") Robot: turn_left(40) & forward(1.3) & turn_right(80) & forward(1.7)	

Table 2: Defined action space of Misty II. **Left:** Action categories and signatures. **Middle:** An example task that illustrates the action execution flow. **Right:** Visual illustrations. (a) basic movement primitives, (b) autonomous navigation to language-specified targets, (c) visual perception within FOV, (d) verbal output and clarification queries, (e) emergency halt.

The world model supports collision detection for obstacle avoidance evaluation. When the robot collides with an obstacle, the system stops the robot at the collision point and immediately marks the task as failed due to safety violation.

Action Space We define navigation, perception, and communication actions based on Misty II's hardware (Table 2; example tasks in Table 1).

- **Navigation:** The four low-level primitives (`forward()`, `backward()`, `turn_left()`, `turn_right()`) provide basic mobility. For spatial navigation, the VLM extracts target objects, estimates relative position, and `navigate(t, d, a)` executes the action sequence.
- **Visual perception:** Misty captures four images at 90-degree intervals for 360-degree coverage. The VLM searches for targets across views; if multiple instances are found, clarification is triggered.
- **Communication:** `speak()` produces verbal output without soliciting input, while `clarify()` generates targeted questions when ambiguity triggers positive friction.

Algorithm 1 formalizes the friction-enabled dialogue loop used in synthetic evaluation, continuing until goal conditions are satisfied or maximum turns reached.

Prompt Design The VLM prompt establishes the robot's identity and action space, defines positive friction behaviors, and specifies a hierarchical ambiguity detection protocol. Inspired by (İnan et al., 2025), ambiguity types are checked in priority order: *contextual safety* (highest, preventing hazardous execution), *trajectory* (avoiding collisions), and *referential & implicit precondition* (establishing common ground). Each type includes a definition and exemplars. The VLM outputs structured JSON specifying action and parameters, which the ActionExecutor parses and dispatches to robot control functions.

4 Experiments and Findings

This section describes the experimental conditions and presents results demonstrating that friction-enabled conversations achieve significantly higher

Algorithm 1 PONDER Dialogue Loop

Require: Scene structure $\mathcal{S}$, Simulated user agent $\mathcal{U}$, task goal $\mathcal{G}$, max turns $T_{\max}$
Ensure: Task evaluation result, interaction log $\mathcal{L}$

$\mathcal{W} \leftarrow$ InitWorldModel($\mathcal{S}, \mathcal{G}$)
$u_1 \leftarrow \mathcal{U}$.GenerateUserUtterance($\mathcal{G}$)
$H \leftarrow \emptyset$ {Conversation history}
$t \leftarrow 1$
while $t \leq T_{\max}$
and not GoalAchieved($\mathcal{W}, \mathcal{G}$) **do**
  $v_t \leftarrow$ GetRobotPOV($\mathcal{W}$)
  $r_t \leftarrow$ VLM(u_t, v_t, H)
  $H \leftarrow H \cup \{(u_t, r_t)\}$
  if r_t[action] = CLARIFY **then**
    Speak(r_t.Q)
    $u_{t+1} \leftarrow \mathcal{U}$.Disambiguate($r_t$.Q, $\mathcal{W}$)
    $t \leftarrow t + 1$
  else
    ExecuteAction(r_t)
    $\mathcal{W} \leftarrow$ UpdateWorldModel($\mathcal{W}, r_t$)
    if not GoalAchieved($\mathcal{W}, \mathcal{G}$) **then**
      $u_{t+1} \leftarrow \mathcal{U}$.FollowUp($\mathcal{W}, \mathcal{G}$)
      $t \leftarrow t + 1$
    end if
  end if
end while
return EvaluateSuccess($\mathcal{W}, \mathcal{G}, H$)

Figure 4: Two top-down views of the setup of two example real-world tasks with Misty being blocked by an obstacle and Misty seeing two identical referents. This is used for the real-world experiments and user study.

task success and goal completion rates compared to frictionless baselines.

We evaluate our architecture through two complementary approaches: real-world user studies on the physical Misty II Robot and large-scale synthetic experiments using simulated dialogues.

4.1 Real-World User Study

Setup For user studies, participants interact with a physical Misty II robot through voice commands. An example task scenario is demonstrated in Figure 4. The system integrates: (1) Google Speech API with wake word detection for speech recognition, (2) GPT-5 nano as unified VLM for spatial reasoning and friction generation, (3) Misty's onboard text-to-speech for verbal responses. Participants complete tasks across ambiguity scenarios in counterbalanced order, with sessions video recorded for post-hoc analysis.

Protocol We recruit 8 participants to interact with the Misty II robot through voice commands. Each participant completes 6 tasks (2 shared for inter-annotator agreement, 4 random for scalability) spanning referential and trajectory ambiguities. Pairwise weighted Cohen's Kappa on satisfaction ratings indicates almost perfect agreement ($\kappa_w = 0.902$). Participants are assigned to both conditions in counterbalanced order:

- **Misty with Friction**: Full system with positive friction enabled. The robot asks clarifying questions when ambiguity is detected.
- **Misty without Friction**: Control condition with friction disabled. The robot executes commands immediately based on its best interpretation.

After each task, participants rate the interaction on 5-point Likert scales (1=very unsatisfied, 5=very satisfied). Sessions are video recorded and task success is determined by whether the robot reached the correct target object or completed the intended action. We have the IRB from our institution for this protocol. For more details of the experimental procedure, please refer to Appendix A.

4.2 Simulated Experiments

Setup To complement user studies with controlled, large-scale experiments, we implement a synthetic dialogue simulation. We employ GPT-4o-mini as the user agent, selected for its balance of strong instruction-following capabilities and API cost efficiency, enabling scalable experimentation across hundreds of trials. We prompt the agent to generate commands following a "vague-to-specific" protocol: initial commands are intentionally under-specified (e.g., "go to the plant"), and the agent provides disambiguation only when the robot asks clarifying questions.

Protocol We run 500 episodes per model configuration using our synthetic dialogue simulation framework comparing four model configurations:

- **Zero-shot VLM**: Stateless processing with no conversation history and no friction. Each command is processed independently. VLM is fed with minimum prompt engineering that only specifies action space and return type
- **Zero-shot Multi-turn**: Same prompt engineering work as Zero-shot VLM but with conversation memory enabled.
- **PONDER without Friction**: Full prompt engineering with conversation history, but friction mechanisms disabled.
- **PONDER**: Complete system with all five friction categories enabled (probing, assumption reveal, overspecify, reflective pause, reinforcement).

All models use GPT-5 nano as the underlying vision-language model, selected for its strong multimodal reasoning capabilities and low inference cost, which enables large-scale evaluation. Task scenarios are randomly sampled from our scenario pool, covering all ambiguity types in Table 1.

4.3 Evaluation Metrics

We evaluate task success by comparing the final world state against goal conditions. The task evaluator parses task goals into verifiable conditions and checks each against the world model state and interaction log. For each task, we measure three types of goal completion conditions as follows.

For *navigation* goals (e.g., "go to the plant"), we compute Euclidean distance between final robot position (x_r, y_r) and target object coordinates (x_t, y_t). Task success requires $d \leq \tau$ where $\tau = 1.0$m by default.

For *perceptual* goals (e.g., "count the chairs"), we analyze the interaction log for robot utterances containing appropriate responses. If the count from robot utterances match the actual count of the objects defined in the world state, task is considered successful.

For *find object* goals (e.g., "find the bag"), if the robot performed 360° scan and provided a verbal description of the target object's location, task is marked successful.

If a task contains multiple goal conditions (e.g., "navigate to desk AND check if laptop is on"), we compute:

$$\text{Goal Completion Rate} = \frac{\text{Conditions Met}}{\text{Total Conditions}} \quad (3)$$

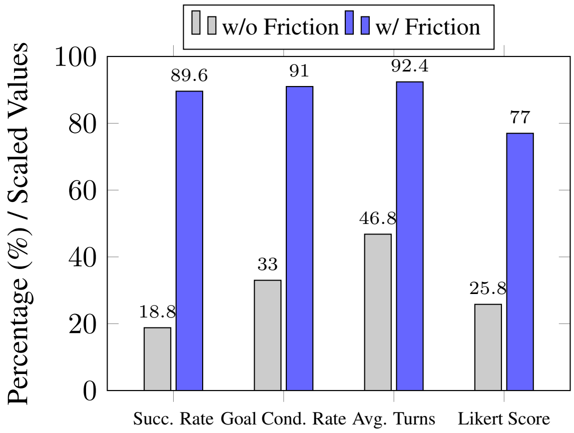


Figure 5: Real-world user study results (N=8 participants). Success Rate: percentage of tasks fully completed; Goal Cond. Rate: average percentage of goal conditions satisfied per task; Avg. Turns: scaled ×40 for visualization (original values: 1.17 and 2.31); Likert Score: scaled from 1-5 to 0-100 (original values: 1.29 and 3.85). Friction-enabled interactions achieve significantly higher performance across all metrics while requiring approximately one more dialogue turn.

A task is marked successful only when *all* goal conditions are satisfied (Goal Completion Rate = 1.0).

Beyond binary success, we also track: (1) *Conversational Turns*: total dialogue exchanges; (2) *Clarifications*: number of friction-induced questions; (3) *Safety Violations*: collision attempts or hazard zone entries;. These metrics enable analysis of the efficiency-accuracy tradeoff inherent in friction-enabled dialogue.

4.4 Findings

Friction-Enabled System Outperforms All Baselines Table 3 presents synthetic experiment results (500 episodes per model). The friction-enabled system achieves 74.8% task success versus 60.25% for the frictionless variant (+14.55pp). Prompt engineering contributes independently: Misty w/o Friction (60.25%) outperforms Zero-shot Multi-turn (44.8%) by 15.45pp. Conversation history alone provides reasonable benefit as Zero-shot Multi-turn (44.8%) exceeds Zero-shot VLM (37.8%) by 7%.

Positive Friction Disambiguates in Both Real-World and Simulated Settings Figure 5 shows consistent patterns in the user study. Friction-enabled conversations achieve 89.6% task success versus 18.8% without friction. Without clarification, the robot frequently selected incorrect referents or ignored obstacles (33.0% goal completion); with friction, ambiguities were resolved before execution (91.0% goal completion). The additional cost of conversational turns is minimal: 2.31 turns on average versus 1.17 turns.

Model	Success Rate (%)	Goal Cond. Rate (%)	Avg. Turns
Zero-shot VLM	37.8	22.0	1.74
Zero-shot Multi-turn	44.8	24.5	1.48
Misty w/o Friction	60.3	47.8	1.38
Misty w/ Friction	**74.8**	**55.8**	1.92

Table 3: Synthetic experiment results (500 episodes per model). Success Rate: percentage of tasks with all goal conditions met; Goal Cond. Rate: average percentage of goal conditions satisfied per task.

Users Prefer Dialogue with Friction Friction-enabled interactions received significantly higher Likert ratings (3.85 vs. 1.29 on a 5-point scale). Participants reported greater trust when the robot clarified before acting.

Real-World vs. Synthetic Comparison Real-world experiments show a larger friction effect (89.6% vs. 18.8%) compared to synthetic experiments (74.8% vs. 60.25%). The lower baseline performance in real-world settings likely reflects the added complexity of noisy physical execution and natural language variation. Despite these challenges, friction-enabled interactions achieve comparable or higher success rates across both settings, demonstrating the robustness of the approach.

4.5 Discussion

Our results provide empirical evidence for Clark's claim that grounding effort reduces total collaborative cost (Clark, 1996a; Clark and Wilkes-Gibbs, 1986). Friction-enabled interactions require approximately one additional turn but dramatically reduce costly failure modes: incorrect navigation, collisions, and user dissatisfaction. The apparent inefficiency of clarification is offset by avoiding physical error recovery, a tradeoff absent in text-based dialogue.

The synthetic experiments reveal complementary factors: conversation history contributes 7%, structured prompt design adds 15.5%, and friction provides 14.5%. The larger performance gap in real-world experiments (70.8% vs. 14.5% synthetic) suggests positive friction is especially valuable when user behavior is unpredictable and real-world noise prevails, precisely the conditions under which deployed systems must operate.

User preference data challenges the assumption that turn minimization should be a primary objective. Likert ratings improved from 1.29 to 3.85 despite additional overhead, consistent with Clark and Wilkes-Gibbs (1986)'s observation that interlocutors willingly invest effort to establish mutual understanding. For embodied systems, users would rather answer a clarifying question than watch a robot navigate incorrectly. These findings suggests that dialogue system design should optimize for collaborative success rather than brevity.

5 Conclusion

We present PONDER, the first implementation of positive friction on physical robotic hardware. In user studies, friction-enabled conversations achieve 89.6% task success versus 18.8% without friction, while synthetic experiments (500 episodes per model) show 74.8% versus 60.3%. Users rate friction-enabled interactions significantly higher (3.85 vs. 1.29), indicating that clarification improves rather than disrupts the experience. Our work demonstrates that deliberately slower conversations produce better outcomes in embodied settings where actions have physical consequences.

Several directions emerge from this work: extending positive friction to diverse robotic platforms such as bimanual manipulators or industrial robots, learning adaptive friction policies through reinforcement learning or user modeling, and combining friction mechanisms with active learning for continual improvement. We release our source code and synthetic dataset to support further research on friction-based embodied dialogue.

Limitations

Our evaluation has three main constraints. First, Misty II's limited action space restricts us to spatial navigation and object localization; future work should deploy positive friction on more diverse platforms to evaluate the mechanism more comprehensively. Second, our user study involves 8 participants in controlled laboratory settings; naturalistic home or workplace environments may reveal different friction patterns and user preferences. Third, the optimal balance between efficiency and friction likely varies across user populations and task contexts, warranting investigation of personalized friction strategies.

6 Ethics Statement

Beyond improving user experience, positive friction serves as a safety mechanism: by requiring explicit confirmation before executing ambiguous commands, the system reduces the risk of collisions, falls, and other physical hazards that arise from premature action execution.

Our study involves two primary ethical considerations: the involvement of human participants in user studies and the use of large language models for friction detection. All human subject recruitment and experimental procedures adhered to our institution's IRB protocols. Additionally, we employ closed-source LLMs (specifically GPT-4o for our experimental pipeline and automatic friction detection, and GPT-4o-mini for the simulated user agent in synthetic experiments), acknowledging that such models may encode biases present in their proprietary training data. However, as our dataset does not contain controversial content or emotionally sensitive topics, we do not anticipate significant bias propagation or amplification from our use of these models.

We also disclose that AI tools were used during the preparation of this manuscript: Nano Banana for generating figures, GPT-5.2 for refining textual clarity, and Claude for assisting in code development. All AI-generated content was reviewed and verified by the authors.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, and 26 others. 2022. Do as i can, not as i say: Grounding language in robotic affordances.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683.

Adithya Bhaskar, Tushar Tomar, Ashutosh Sathe, and Sunita Sarawagi. 2023. Benchmarking and improving text-to-SQL generation under ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7053–7074, Singapore. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. pages 8718–8735.

Erdem Bıyık, Nicolas Huynh, Mykel J Kochenderfer, and Dorsa Sadigh. 2024. Active preference-based gaussian process regression for reward learning and optimization. volume 43, pages 665–684. SAGE Publications Sage UK: London, England.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, and 5 others. 2024. π_0: A vision-language-action flow model for general robot control.

Joyce Y. Chai, Rui Fang, Changsong Liu, and Lanbo She. 2017. Collaborative language grounding toward situated human-robot dialogue. *AI Magazine*, 37(4):32–45.

Joyce Y. Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. 2014. Collaborative effort towards common ground in situated human-robot dialogue. In *ACM/IEEE International Conference on Human-Robot Interaction, HRI'14, Bielefeld, Germany, March 3-6, 2014*, pages 33–40. ACM.

Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee Wong. 2024. Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9796–9810.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. 2023. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*.

Herbert H Clark. 1996a. *Using language*. Cambridge university press.

Herbert H Clark. 1996b. *Using Language*. Cambridge University Press.

Herbert H Clark and Susan E Brennan. 1991a. Grounding in communication.

Herbert H Clark and Susan E Brennan. 1991b. Grounding in communication. In *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association.

Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, and 275 others. 2025. Open x-embodiment: Robotic learning datasets and rt-x models. *Preprint*, arXiv:2310.08864.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10602–10621, Singapore. Association for Computational Linguistics.

Vardhan Dongre, Xiaocheng Yang, Emre Can Acikgoz, Suvodip Dey, Gokhan Tur, and Dilek Hakkani-Tur. 2025. ReSpAct: Harmonizing reasoning, speaking, and acting towards building large language model-based conversational AI agents. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 72–102, Bilbao, Spain. Association for Computational Linguistics.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Haimin Hu, David Isele, Sangjae Bae, and Jaime F Fisac. 2024. Active uncertainty reduction for safe and efficient interaction planning: A shielding-aware dual control approach. *The International Journal of Robotics Research*, 43(9):1382–1408.

Wenlong Huang, Fei Xia, Dhruv Shah, Danny Driess, Andy Zeng, Yao Lu, Pete Florence, Igor Mordatch, Sergey Levine, Karol Hausman, and 1 others. 2023a. Grounded decoding: Guiding text generation with grounded models for embodied agents. *Advances in Neural Information Processing Systems*, 36:59636–59661.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Tomas Jackson, Noah Brown, Linda Luu, Sergey Levine, Karol Hausman, and brian ichter. 2023b. Inner monologue: Embodied reasoning through planning with language models. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 1769–1782. PMLR.

Mert Inan, Anthony Sicilia, Alex Xie, Saujas Vaduguru, Daniel Fried, and Malihe Alikhani. 2025. Identifying & interactively refining ambiguous user goals for data visualization code generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25257–25274.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. 2023. VIMA: Robot manipulation with multimodal prompts. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 14975–15022. PMLR.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. 2025. Openvla: An open-source vision-language-action model. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 2679–2713. PMLR.

Przemyslaw A. Lasota, Terrence Fong, and Julie A. Shah. 2017. A survey of methods for safe human-robot interaction. *Foundations and Trends® in Robotics*, 5(4):261–349.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. 2023. Code as policies: Language model programs for embodied control. *Preprint*, arXiv:2209.07753.

Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. 2023. Text2motion: from natural language instructions to feasible plans. *Autonomous Robots*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Timothy Obiso, Kenneth Lai, Abhijnan Nath, Nikhil Krishnaswamy, and James Pustejovsky. 2025. Dynamic epistemic friction in dialogue. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 323–333, Vienna, Austria. Association for Computational Linguistics.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, and Dilek Hakkani-Tur Gokhan Tur and. 2022. TEACh: Task-driven Embodied Agents that Chat. In *Conference on Artificial Intelligence (AAAI)*.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83, Ann Arbor, Michigan. Association for Computational Linguistics.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 627–635, Fort Lauderdale, FL, USA. PMLR.

Dorsa Sadigh, Anca Dragan, Shankar Sastry, and Sanjit Seshia. 2017. Active preference-based learning of reward functions.

Rupak Sarkar, Neha Srikanth, Taylor Pellegrin, Rachel Rudinger, Claire Bonial, and Philip Resnik. 2025. Understanding common ground misalignment in goal-oriented dialog: A case-study with Ubuntu chat logs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3200–3215, Vienna, Austria. Association for Computational Linguistics.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. 2022. Cliport: What and where pathways for robotic manipulation. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 894–906. PMLR.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Tejas Srinivasan and Jesse Thomason. 2025. Adjust for trust: Mitigating trust-induced inappropriate reliance on ai assistance. *Preprint*, arXiv:2502.13321.

Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:25–55. Publisher Copyright: Copyright © 2020 by Annual Reviews. All rights reserve.

David R Traum. 1994. A computational theory of grounding in natural language conversation. *PhD thesis, University of Rochester*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Michael JQ Zhang, W Bradley Knox, and Eunsol Choi. 2024a. Modeling future conversation turns to teach llms to ask clarifying questions. *arXiv preprint arXiv:2410.13788*.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. 2024b. CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10746–10766, Bangkok, Thailand. Association for Computational Linguistics.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, and 35 others. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR.

Mert İnan, Anthony Sicilia, Suvodip Dey, Vardhan Dongre, Tejas Srinivasan, Jesse Thomason, Gökhan Tür, Dilek Hakkani-Tür, and Malihe Alikhani. 2025. Better slow than sorry: Introducing positive friction for reliable dialogue systems. *Preprint*, arXiv:2501.17348.

A User Study Protocol

A.1 Instruction Script

An example of the instructions we provide to the participants has been shown in Figure 7.

A.2 Task Prompts

Participants were given high-level goal descriptions without specifying how to phrase commands:

- *Referential*: "Final goal: Misty reaches the target plant" (two identical plants visible)
- *Trajectory*: "Final goal: Misty reaches the trash bin" (obstacle blocking direct path)

A.3 Post-Task Questionnaire

After each task, participants responded to: *"How satisfied are you with this interaction?"* on a 5-point Likert scale (1=Very Unsatisfied, 5=Very Satisfied).

B Example Interaction Log

Figure 6 shows a representative multi-turn interaction demonstrating trajectory ambiguity resolution through probing friction.

```
================================================================================
Misty Conversation Log - 2025-12-17 19:13:39
================================================================================

Turn 1:
--------------------------------------------------------------------------------
User: go near the trash bin

VLM_raw_output: {
  "friction_type": "probing",
  "action": "clarify",
  "distance": 0,
  "text": "I see the trash bin ahead, but there's a small case on the floor blocking the direct path.
  Should I go around it or stop nearby?",
  "target_object": "trash bin",
  "turn_degrees": 0,
  "clarification_needed": "Should I go around the obstacle (the case) or stop near the trash bin?",
  "confidence": "medium"
}

Misty:
  1. OK, let me think about that
  2. I see the trash bin ahead, but there's a small case on the floor blocking the direct path.
  Should I go around it or stop nearby?

Turn 2:
--------------------------------------------------------------------------------
User: you should go around it

VLM_raw_output: {
  "friction_type": "assumption_reveal",
  "text": "I'll navigate around the obstacle on the left side where there's more clearance.",
  "confidence": "high",
  "actions": [
    {"action": "turn_left", "turn_degrees": 30, "distance": 0},
    {"action": "forward", "distance": 1.5},
    {"action": "turn_right", "turn_degrees": 40, "distance": 0},
    {"action": "forward", "distance": 0.9}
  ]
}

Misty:
  1. OK, let me think about that
  2. I'll navigate around the obstacle on the left side where there's more clearance.
  3. Turning left 30 degrees
  4. Moving forward 1.5 meters
  5. Turning right 40 degrees
  6. Moving forward 0.9 meters

================================================================================
Session ended - 2025-12-17 19:15:25
Total turns: 2
================================================================================
```

Figure 6: Example interaction log demonstrating positive friction dialogue.

Participant Instructions: Voice-Controlled Robot Task Completion

Overview

In this study, you'll work with Misty, a voice-controlled robot that can navigate environments and interact with objects. Your goal is to communicate naturally with Misty to help it complete certain tasks.

Misty uses VLM to understand your commands and may ask you clarifying questions when it needs more information to successfully complete tasks. This is a normal part of the interaction. Please respond to Misty's questions naturally, just as you would in a conversation with another person.

How to Interact with Misty

Starting a conversation:

- Say "Hey Misty" to get the robot's attention
- Wait for a visual cue (LED color change) indicating Misty is listening
- Speak your command clearly in natural language

During the conversation:

- Misty will respond through its speaker
- If Misty asks clarifying questions, answer them naturally
- Always say “Hey Misty” first for it to start listening
- You can give commands in your own words. There's no need to use specific phrases

Your Task

Final goal: Misty reaches the target plant

Take your time and communicate with Misty as you would with a helpful assistant. There are no wrong questions to ask or wrong ways to phrase your instructions.

Figure 7: Participant Instruction