

Looking for Objects and Looking for Help: Robot Gaze Elicits Human Cooperation

Anna Manaseryan
annamanaseryan@u.
boisestate.edu
Computer Science
Boise State University

Thi Lan Anh Bui
anhbui@u.
boisestate.edu
Computer Science
Boise State University

Casey Kennington
caseykennington@
boisestate.edu
Computer Science
Boise State University

Abstract

Robots that search for objects on their own must decide what to do when the search fails, and how they signal that failure to a human partner can determine whether the interaction turns into collaboration. Prior work on human–robot interaction has studied nonverbal help-seeking and social gaze separately, but rarely compares gaze against a matched verbal failure signal inside a live, real-time embodied search task. We present an embodied dialogue system, built on the Retico incremental processing framework, where the Cozmo robot searches for a specified object and signals failure either verbally or through gaze towards the human partner. We first compare a rule-based planner against a vision-language model (VLM) planner that reasons over scene semantics, finding that the VLM planner succeeds far more often (87.2% vs. 65.9%). We then test whether gaze can close this gap without adding reasoning capability to the rule-based planner. Adding gaze as the stuck signal raises success to 87.87%, matching the VLM planner, while a verbal stuck signal only raises it slightly (68.75%). Results further show that this effect is largely implicit: participants helped more in the gaze condition without consistently reporting that they felt invited to do so. Our findings suggest that gaze can substitute for reasoning capability a robot lacks, helping it elicit assistance from a human partner.

1 Introduction

Multimodal, situated interaction is central to natural human–robot dialogue where people and robots coordinate not only through words, but also through gestures, emotion, and other information (McNeill and Kennington, 2020; Admoni and Scassellati, 2017; Novikova et al., 2015; Mutlu et al., 2012). In a setting where a robot explores a space for objects, the robot must signal a need for help to the human when the exploration fails, otherwise the human collaborator cannot know whether some

kind of an intervention to help the robot will actually succeed. When a robot fails to find an object, the way it signals that failure matters. A verbal statement such as *I cannot find the object* communicates the need for help, but does not naturally open the floor for collaboration.

Prior work shows that nonverbal signals can act as requests for help during collaborative tasks (Cha and Matarić, 2016). One specific nonverbal signal — turning to face the person and holding the person’s gaze — is a social act that can implicitly invite participation without requiring an explicit request (Mutlu et al., 2009; Cha and Matarić, 2016). *Eye gaze* is an important communicative modality that can invite participation from humans by shaping conversational footing (Mutlu et al., 2009), and other work shows that eye contact changes how later robot gaze is processed (Kompatsiari et al., 2018). It has also been shown that robot behaviors display emotion (Novikova et al., 2015; McNeill and Kennington, 2019) and that emotional expression encourages helping in collaborative tasks between robots and humans (McNeill and Kennington, 2020; Zhou and Tian, 2020). Taken together, these findings suggest that failure-triggered gaze toward a human partner coupled with an appropriate emotional behavior could elicit collaboration in a live object-search task. Following work on nonverbal help-seeking in human–robot collaboration (Cha and Matarić, 2016) and on social eye gaze in human–robot interaction (HRI) (Admoni and Scassellati, 2017; Mutlu et al., 2009), in this work, we task a human and a robot to collaboratively seek out objects, but when a robot is ‘stuck’ it uses eye gaze as a communicative act; i.e., at the moment the robot cannot find a target, a policy must select *how* to disclose failure, which is then realized as either speech or embodied orientation with gaze toward the human.

Our work uses a system built on the Retico incremental processing and ‘robot-ready’ spoken di-

dialogue system frameworks (Michael, 2020; Manaseryan et al., 2025). Our Experiment first establishes a baseline: Human participants tell the robot which object they want it to find and we use a language model to autonomously control a robot to explore objects in a confined space, which shows that a language model can generate plans better than a rule-based planner, but in both systems there are many cases where the robot is unable to find a specified object. That leads to our primary experiment where we ask: *How can the robot elicit help from the human when the robot cannot find the object?* In that experiment, two conditions share mid-search frustration displays; only the stuck signal differs: either a verbal ‘give-up’ (“can’t find”) or a physical turn toward the participant with held gaze. Thus emotion is held constant across conditions and cannot explain the between-condition difference; the contrast is the stuck cue itself. We find that the robot is able to complete more tasks (i.e., find more objects) when it is able to elicit help from the user both using speech and gaze, but more so using gaze, to the degree that the interactive help from human participants parities a more sophisticated planner used to seek objects.

Our contributions are as follows:

1. A real-time embodied dialogue system built on Retico (Michael, 2020; Manaseryan et al., 2025) that integrates automatic speech recognition, object detection and search planning, generative multimodal emotional behavior, and robot execution in a single end-to-end interactive loop.
2. A controlled comparison of visual planning strategies for physical object search task on Cozmo, establishing the live search setting on which Cozmo requests help when it cannot find the object.
3. A controlled comparison of gaze after failure versus a matched verbal give-up. We evaluate spontaneous collaborative guidance and how participants interpret the stuck cue as an invitation to help. We also observe whether directional hints after the stuck signal allow the robot to continue searches that would otherwise end.

In the following section we compare our work to others, then explain our methodology. We then explain two experiments where we (1) evaluate alternative planners for our object-search task, then

(2) test whether gaze at task failure invites collaboration relative to a verbal baseline.. We then offer limitations and conclude.

2 Related Work

Social eye gaze is a communicative channel in HRI. Robots use gaze to manage attention, show engagement, and shape who is invited to take part (Admoni and Scassellati, 2017). Mutlu et al. (2009) showed that gaze cues alone can assign conversational footing (addressee, bystander, or overhearer), with participants following robot-assigned roles at high rates. Later work extended these ideas to turn-taking and dialogue structure with more than two people (Mutlu et al., 2012). Other evidence shows that looking at a person’s face before a later gaze shift changes how that cue is processed (Kompatsiari et al., 2018), and that human gaze toward a robot often comes before joint action in collaborative assembly (Lavit Nicora et al., 2024). We build on this by treating the *robot’s* gaze toward the human after failure as an invitation to collaborate once search has stalled.

A parallel line of work asks how robots should request help when collaboration breaks down. Cha and Matarić (2016) found that nonverbal help-request signals (light and sound at different urgency levels) change both helping behavior and how people see the collaboration for a non-humanoid robot. Other related work studies the reverse case. *Human* gaze can show that the person has noticed robot failure, including real-time detection from gaze patterns (Tabatabaei et al., 2025a,b). This work (specifically, when robot requests help when it cannot find the object) instead treats *robot-to-human* gaze itself as the failure signal.

Visual planning for physical robot tasks has moved from rule-based detection toward vision-language models that reason over scene semantics rather than geometry alone (Driess et al., 2023). Recent VLMs such as Qwen2.5-VL-7B (Bai et al., 2025) can perform object localization and spatial reasoning from raw scene context, without relying on bounding-box confidence thresholds. Experiment 1 compares this approach directly against a YOLO-based rule-based planner on a live object-search task.

Related to eye gaze, affect and emotion form a separate channel for eliciting help from humans. Emotionally expressive robots are more likely to receive collaboration in joint tasks, including help

from the human and task success time (McNeill and Kennington, 2020; Zhou and Tian, 2020). A specific emotion, *frustration* behaviors are also argued to matter for successful human–robot collaboration more broadly (Weidemann and Rußwinkel, 2021). In this work, we use emotional behaviors to improve the interactive qualities of our robot, though we do not compare an emotional vs. non-emotional robot display as the emotion (e.g., *frustration*). Emotional displays are a shared baseline that signals something is wrong, but not specifically that help is requested.



Figure 1: The Cozmo robot (Digital Dream Labs), the embodied platform used in both experiments.

3 Method

Robot Platform Both of our experiments use Cozmo, a small toy-like robot licensed by Digital Dream Labs (see Figure 1). Cozmo has a front-facing RGB camera (320×240 at 15 frames per second), an OLED screen for facial expressions, a movable head, a lift arm, and drive wheels for tabletop navigation. Control is made through the Cozmo Python SDK. These capabilities make it a suitable platform for multimodal face-to-face interaction and for our task: the camera supports vision-based search, the screen and body actuators support emotional behaviors, and the head and drive motors support physical orientation toward a human partner. It has also been shown in prior work that a non-verbal Cozmo elicited cooperation from humans based on how its behaviors displayed emotion (McNeill and Kennington, 2020), and it has also been shown that people view Cozmo as a child-level robot that is 2-3 years old in comparison with human cognition (Plane et al., 2018), which is important because children often use physical displays of emotion (e.g., frustration) and eye gaze as a signal of confusion and to elicit help from others before they are able to speak.

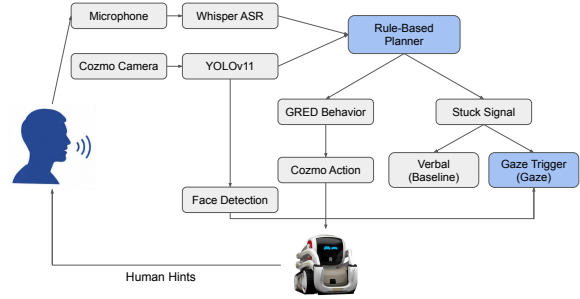


Figure 2: System architecture for our experiment. The rule-based planner receives both ASR and YOLO streams and coordinates all robot actions, including GRED emotional behaviors and the stuck signal (verbal ‘stuck’ state; gaze orientation in Gaze condition).

System Architecture Figure 2 shows the full system. Our system is built on the Retico incremental processing framework (Michael, 2020; Manaseryan et al., 2025), which connects modules by passing Incremental Units (IUs) as fine-grained units of information between modules (Schlangen and Skantze, 2011). Such fine-grained, incremental processing is an important property of natural dialogue between robots and people (Manaseryan et al., 2025). Two input modalities feed the central planner in parallel: (1) the speech modality, which uses a Microphone Module to stream raw audio to Whisper ASR Module, which emits word-level updates as speech arrives and a committed utterance once the participant finishes speaking, and (2) the vision modality, which uses a Cozmo Camera Module to read image frames from Cozmo’s front camera and pass them to a YOLOv11¹ Module, which emits detected object classes, confidences, and bounding boxes. The planner receives both streams (transcribed speech and bounding box / object labels), maintains search state, and emits action commands to Cozmo Action Module, which serializes them in a queue and executes them one at a time through the Cozmo Behaviors Module. Actions include search movements, emotional behaviors, and spoken output. A separate Face Detection Module acts as another modality input that scans for the participant’s face at startup and stores their location. In the experiment, this module is also used to enable Cozmo to angle its face towards and gaze at the human participant when needed.

¹<https://github.com/ultralytics/ultralytics>

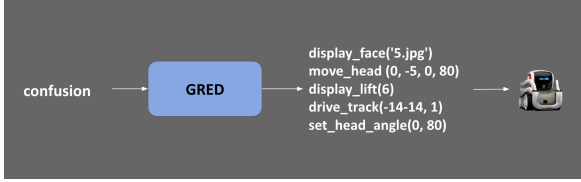


Figure 3: GRED maps an emotion label (here, confusion) to a sequence of Cozmo actions, including facial display, head motion, lift, and drive, which are then executed on the robot.

GRED Emotional robot behavior is produced by GRED (Generate Robot Emotional Displays) (Baral et al., 2025), a generative model built on GPT-2 Medium. GRED takes an emotion label and produces a novel sequence of parameterized Cozmo actions, including facial displays on an OLED screen, short vocalizations, driving patterns, head angles, and lift-arm movements, such that playing back those actions on the robot would make an observer think the robot is displaying the specified emotion. The model was fine-tuned on hand-designed behavior sequences across six grouped emotion categories, including Anger/Frustration and Joy/Hope. Each behavior is a sequence of symbolic action tokens (for example, `display_oled_face_image-face2 turn_in_place 30 230`). An example of input and output can be seen in Figure 3. At inference time, our search Planner module emits an emotion label to GRED and executes the returned behavior on Cozmo. In this work, the Planner Module uses `anger_frustration` during mid-search escalation when the target is still missing, and `joy_hope` when the object is found. In our experiments, both conditions share these GRED displays. Displaying emotions is very important as the robot needs to have a basic ability to display emotional behaviors or the evaluation would not be convincing as the robot would be perceived by humans as “too robotic.” It was shown in prior work that adding emotional qualities to the robot increased interaction and desire from human participants to help the robot in learning tasks (McNeill and Kennington, 2020).

Planner When a participant utters a command to find a specific object, the planner enters a searching state where it takes random turns and drive distances while YOLOv11 checks each incoming camera frame for the target object. If the target is not detected after a specific time (20 seconds for our experiments), the robot displays frus-

tration using the GRED Module. After 10 more seconds without a detection, the search reaches a stuck state. To establish the base capacity of the system, we compare two planners: a rule-based planner and a visual language model (VLM) based planner. Both planners use the YOLOv11 bounding box and object label information. The rule-based planner follows a set of rules to make random movements to explore the space until the specified object is found, and moves closer to the object once it is. The VLM-based planner uses a local LM to make a plan for robot actions (i.e., non-random) to move closer to a specified object.

The final experiment extends the Planner with a stuck state and hint state. After the stuck state is reached, the Planner enters the hint state and listens for speech from the participant. If a hint from the participant is given (e.g., *turn right* or *look behind you*), Cozmo executes the movement and checks YOLO for the target. If three consecutive hints do not result in the robot finding the object, the system enters the stuck state again. If the target is detected at any point during search or hint states, the Planner enters an idle state and Cozmo says *found it* while GRED generates and plays a `joy_hope` display on Cozmo. How the stuck signal is delivered differs by either a speech event where the robot utters *can’t find*, or a *gaze trigger*, explained below.

Gaze Trigger In the gaze condition, the stuck state triggers a physical act rather than speech. Because Cozmo’s screen eyes have limited expressivity, the cue is not eye movement alone: Cozmo rotates its whole body to face the participant, adjusts its head toward their face, and holds that facing orientation. This is made possible by the Face Detection Module, which rotates in place at startup until it detects the participant’s face and stores that location in Cozmo’s SLAM Module. When the stuck state triggers, Cozmo uses that stored position to turn toward the person and sets its head angle by distance so it is oriented at their face. Cozmo then holds that orientation for three seconds so the participant can register the facing cue. This pause is deliberate: prior work shows that establishing direct eye contact before any subsequent gaze shift changes how that cue is processed (Kompatsiari et al., 2018). After the hold, Cozmo remains facing the participant and does not turn back immediately. The head resets only when the next action begins, whether that is a hint movement or a new

search command from the participant, so the signal stays visible until the participant responds. The baseline condition produces no physical movement and no orientation change at the stuck moment, so the between-condition contrast is this embodied facing/gaze cue versus the verbal give-up.

4 Experiment 1: Planner Comparison

We first establish the base capability of the system given our task which requires tabletop robot navigation: a human participant tells the Cozmo robot to find a specific object type, then Cozmo autonomously searching for the targeted object on a tabletop under two visual planners that share the same robot platform, camera input, safety guards and completion criteria, and differ only in how each frame is interpreted to select the next action. Figure 4 shows the full system.

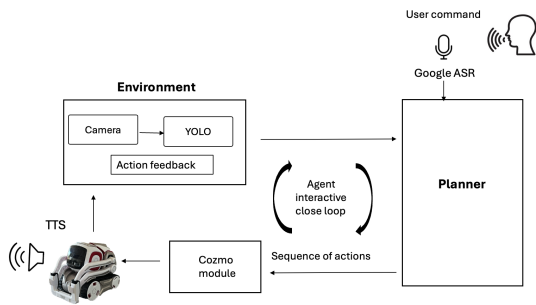


Figure 4: Base system architecture.

Planners. We first evaluate two different Planner variants. The two Planners differ only in how each frame is interpreted to select the next action. The **VLM-planner (Qwen2.5-VL-7B)** interprets the scene to perform object localization, spatial reasoning, and visibility assessment, and selects actions based on semantic scene understanding rather than geometry alone. The **rule-based planner** detects the target object in each frame and maps bounding-box geometry to navigation decisions through rule-based logic.

Procedure: We conducted a within-subjects study with ten participants, with Planner order randomized across participants. Each participant completed two 15-minute sessions, one for each Planner, in the same tabletop workspace with a fixed object set (car, bicycle, chair, and bottle). During each session, participants issued voice commands (e.g., *find the chair*) specifying the target object and observed Cozmo’s behavior. Participants were al-

lowed to reposition Cozmo and the objects between attempts.

After each session (a session could have several instances of the participant instructing Cozmo to find a specific object), participants rated the Planner along two dimensions: task performance (six custom 5-point Likert items assessing whether the robot’s behavior matched the requested task) and perceived intelligence from the the Godspeed sub-scale (Bartneck et al., 2009), consisting of five 5-point semantic differential items.

Metrics. Planner performance is evaluated using objective and subjective measures. Objective measures, drawn from system logs, capture task success rate (whether the system found the object). Subjective measures from post-session questionnaires capture perceived performance through task performance and perceived intelligence ratings. Because participants set their own object arrangements on the table, some failures may result from scene difficulty rather than planner limitations. Subjective ratings therefore complement objective metrics by capturing aspects of behavior not reflected in completion outcomes. Statistical analyses follow the paired within-subjects design and small sample size. Participant-level comparisons use the Wilcoxon signed-rank test.

Results. The VLM planner found the target object in 41 of 47 attempts (87.2%), compared to 27 of 41 for rule-based (65.9%). This difference was significant (Wilcoxon signed-rank test, $p = 0.039$, $r = 0.79$), indicate that the VLM planner is more reliable at finding the target object than rule-based planner. Table 1 shows the detail results.

Table 1: Object-finding performance by system.

System	Found	Not found	Found rate
VLM planner	41 / 47	6 / 47	87.2%
Rule-based planner	27 / 41	14 / 41	65.9%

Subjective ratings followed the same pattern as objective performance. Participants rated the VLM planner higher on Task Performance (4.20 vs. 3.37; Wilcoxon signed-rank test, $p = 0.016$, $r = 0.76$) and Perceived Intelligence (4.02 vs. 3.30; $p = 0.016$, $r = 0.76$), with large effect sizes for both measures. (Appendix D, Figure 9 summarizes the ratings).

Discussion. The VLM-planner outperforms the rule-based planner on both objective and objec-

tive measures, completing 87.2% of tasks versus 65.9% for the rule-based planner, with significantly higher ratings on task performance and perceived intelligence. In most of these cases in which the VLM planner found the target while the rule-based planner did not, participants created complex setups where the target was occluded or in an unusual pose (Figure 5), degrading YOLOv11’s detection confidence. When detection confidence drops, the rule-based planner lacks sufficient information to act, since its decisions depend entirely on bounding-box features. The VLM planner remains effective in these cases by identifying the target from scene-level context.

Specifically, 34% of rule-based planner attempts ended without finding the target, so failure was common even though finding the target was still more frequent. This failure moment is what our experiment is built on: when the robot gets stuck, how should it signal that to the human?



Figure 5: Occluded object and unusual object pose case (images are extracted from Cozmo’s camera view)

5 Experiment 2: Gaze vs. Verbal Stuck Signal

With our planner baseline established, we now turn to the main experiment. Our goal is to evaluate a system that uses the rule-based planner with the added capability of requesting help from the user through eye gaze and compare it to the VLM planner. Our hypothesis is that the rule-based planner, while not as effective as the VLM-planner on its own, will be as or more effective at the object finding task when it elicits help from humans using eye gaze.

Procedure. Our experiment uses a within-subjects design with two conditions: Baseline (ver-

bally uttering *can’t find*) and eye gaze (physical orientation toward the participant). Each participant completes both conditions in randomized order to control for order effects. Both sessions use the rule-based planner. The target objects are the same common household items as in the preliminary experiment (cup, chair, potted plant, bicycle) for which bounding-box detection is reliable. Figure 6 shows the tabletop workspace and a participant observing Cozmo during a session. Before each session, the experimenter places the target object out of Cozmo’s initial camera field of view so the stuck state triggers reliably.

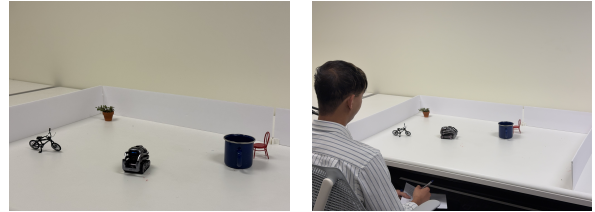


Figure 6: Left: tabletop workspace with Cozmo and target objects (cup, chair, potted plant, bicycle). Right: participant observing Cozmo, holding the paper sheet used to record found objects.

Participants are introduced to Cozmo with the following instructions: “In this session, you will be working alongside a Cozmo robot. Cozmo’s goal is to explore the environment and find as many target objects as possible. Cozmo understands spoken English. If at some point it seems like it needs your help, you can try talking to it.” No hint vocabulary is disclosed and no mention is made of the gaze condition (full instruction sheet in Appendix A). Participants are given a paper sheet to fill in during the session, listing which objects the robot found on its own and which it found with their help (Appendix B). As before, participants can set the objects where they please on the table. Each session covers four objects in fixed order and lasts approximately 20 minutes, followed by a post-session questionnaire containing custom Likert items assessing task performance and collaboration perception, and the Godspeed robot evaluation questionnaire (Bartneck et al., 2009). Total participant time is around one hour. Participants received a \$12 Amazon gift card as compensation. Eight participants took part in the study.

Metrics & Baseline. Only the stuck state differs between conditions. *Baseline* says “can’t find”; *Gaze* turns Cozmo’s body to face the participant

and holds that orientation (a full-body facing cue, not eye motion alone). Mid-search timing and GRED anger_frustration displays are identical in both conditions, as are hint commands after the stuck cue; emotion is therefore not confounded with condition. The manipulated factor is only how failure is disclosed at the stuck moment. The primary measure is objects found with participant help, recorded on a paper notes sheet (Appendix B). Objects found alone and total objects are secondary measures.

After each session, participants completed a questionnaire (Appendix C) with eight custom 5-point Likert items on task experience and collaboration, and the Godspeed questionnaire (Bartneck et al., 2009) covering Anthropomorphism, Animacy, Perceived Intelligence, and Likability. The two custom items target our research question (i.e., *I felt invited to help* and *robot movements made sense when stuck*). All comparisons use the Wilcoxon signed-rank test with effect size r (exact two-tailed p , effect size $r = |z|/\sqrt{N_{\text{nontied}}}$).

Results. Behavioral outcomes: Participants recorded which objects Cozmo found on its own and which it found with their help on a paper sheet during each session.

P	Baseline				Gaze			
	Alone	Help	Total	Miss	Alone	Help	Total	Miss
1	3	1	4	0	3	1	4	0
2	3	1	4	0	2	2	4	0
3	3	0	3	1	3	1	4	0
4	3	0	3	1	2	0	2	2
5	2	0	2	2	3	1	4	0
6	4	0	4	0	3	1	4	0
7	1	0	1	3	2	1	3	1
8	1	0	1	3	2	2	4	0
Sum	20	2	22	10	20	9	29	3
Mean	2.50	0.25	2.75	1.25	2.50	1.13	3.63	0.38
SD	1.07	0.46	1.28	1.28	0.53	0.64	0.74	0.74

Table 2: Object counts per participant. Miss = objects never found in either condition. Help: $W = 0$, $p = 0.031$, $r = 0.899$.

Per-participant counts for found objects are in Table 2. Only 2 of 8 participants helped in Baseline; 7 of 8 helped in Gaze. Objects found with help: Gaze mean = 1.13 (SD = 0.64) vs. Baseline mean = 0.25 (SD = 0.46), $W = 0$, $p = 0.031$, $r = 0.899$. Objects found alone were the same in both conditions (mean = 2.50; $W = 10.5$, $p = 1.0$), showing the effect is specific to the collaboration channel. Total objects trended toward Gaze (3.63

vs. 2.75; $W = 1.5$, $p = 0.188$, $r = 0.724$) but did not reach significance at $N = 8$.

Qualitatively, the conditions looked different at the stuck moment. In Baseline, most participants watched after hearing “can’t find” and did not speak; only two participants offered help across the whole condition, and helping was rare (2 objects found with help in total). In Gaze, after Cozmo turned to face them, seven of eight participants responded with spoken guidance, typically short directional hints such as *turn left*, *turn right*, or *look behind you*, which the planner could execute in the hint state. Objects found alone stayed the same across conditions (20 vs. 20), so the extra successes came from this help channel rather than from better solo search. We did not run a formal coding of every utterance; this account summarizes the pattern already visible in the per-participant counts in Table 2.

Gaze vs. VLM Planner: In the baseline setting, the robot was able to find 20 objects without help and an additional 2 with help (i.e., participants offered hints when the robot said *I’m stuck*). In the Gaze system where the robot looked at the participant when stuck, the robot was able to find 20 objects without help, and an additional 9 objects with help. The baseline system had 68.75% (22 out of 32) successfully found objects, whereas in the Gaze system, 87.87% (29 out of 33) resulted in found objects. Compared to the VLM Planner which had a 87.2% success rate, using the more basic planner with increased interaction using Gaze to elicit help enabled the robot to successfully find objects at a rate on-par (in fact, slightly better) than the vision-language model based planner. This supports our hypothesis that gaze with a simpler planner is comparable to a more sophisticated planner. This is a very positive result, as it shows that simple planning with elicited help from humans, which is akin to the behavior of children who seemingly randomly explore their environment yet look at caregivers for help when frustrated or confused, is an effective and collaborative method of exploration.

Custom Likert items. After each session, participants rated eight custom Likert items (1–5) on task experience and collaboration. Two questions are specific to our task: (Question 6: *I felt invited to participate when the robot seemed unsure what to do*) and (Question 9: *The robot’s movements made sense for when it was unable to find an object*).

Item	Baseline		Gaze		<i>W</i>	<i>p</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
Q3 Interesting	3.88	0.99	3.75	0.71	2.0	1.000
Q4 More time	4.00	1.20	4.13	1.36	2.0	1.000
Q5 Found object [†]	3.13	0.83	3.71	0.76	2.0	0.250
Q6 Felt invited*	3.63	1.19	3.38	1.06	3.5	0.750
Q7 Engaged	3.63	0.92	3.75	0.71	6.0	1.000
Q8 Movements/search [†]	4.00	0.76	3.71	0.76	2.5	0.625
Q9 Movements/stuck*	3.88	0.99	3.63	1.06	1.5	0.750
Q10 Collaboration	3.75	0.71	3.50	1.07	1.5	0.750

Table 3: Custom Likert means (1–5) and exact Wilcoxon statistics. **Bold** rows are primary items (*). [†] $N = 7$ (one missing response). No item reached significance ($\alpha = 0.05$).

No item reached significance (Table 3). Q6: Baseline = 3.63 (SD = 1.19), Gaze = 3.38 (SD = 1.06), $W = 3.5$, $p = 0.750$, $r = 0.274$. Q9: Baseline = 3.88 (SD = 0.99), Gaze = 3.63 (SD = 1.06), $W = 1.5$, $p = 0.750$, $r = 0.463$. Perceived collaboration (Q10) showed no difference either (Baseline = 3.75, Gaze = 3.50; $W = 1.5$, $p = 0.750$).

Two factors may explain these Likert results. First, gaze appears to have worked as an *implicit* cue rather than an explicit one: participants helped more in the gaze condition, which resulted in more objects found, yet did not rate themselves as feeling more invited or as better understanding the stuck signal. We conjecture that a facing/gaze cue can shift collaborative behavior without the observer consciously registering it as a request (Kompatsiari et al., 2018). Second, individual variation was high on both items (SD 1.06–1.19 on a 5-point scale), which reduces power at $N = 8$. Effect sizes ($r = 0.27$ for Q6, $r = 0.46$ for Q9) suggest a larger study would be needed to confirm these trends. See Appendix for Godspeed questionnaire results, which similarly showed no difference in human perceptions between the two systems.

The low helping rate in the verbal baseline also needs explanation. The phrase “can’t find” names failure, but it does not clearly invite participation; participants may have heard it as the end of a trial rather than a request for help. Robot speech can also be hard to hear or parse, so some people may simply have missed or underweighted the cue. We therefore cannot fully separate “gaze/orientation is strong” from “this particular verbal phrase is weak.” A stronger verbal topline would use an explicit invitation (e.g., “can you help me?” or “please tell me where to look”), as in prior work on robot help-

seeking (Cha and Matarić, 2016). We did not run that condition here.

6 Limitations

This study has several limitations. First, both experiments were conducted with relatively small sample sizes (preliminary experiment: $n = 10$, main experiment: $n = 8$), which may limit statistical power, so results may reflect random variation rather than true differences; future studies should replicate these findings with larger, more diverse participant pools. Second, the GRED emotional expressions were not always clear for the participants to interpret, which likely weakened their usefulness as aid cues; we also did not run an emotion-off condition, so we cannot claim that emotion plus gaze is jointly necessary. This limitation could be mitigated by refining the GRED output for more consistent interpretation and by testing an emotion-off ablation in future work. Third, the setup in our experiment was simpler and did not include many failure cases where the baseline system struggled with object detection. Future studies should test more complex and varied environments. Additionally, our experiment evaluated only the rule-based planner, which has limited ability to interpret human-provided hints (e.g. “the cup is behind you”); future work could improve its language grounding or extend the evaluation to VLM-based planners. Finally, both experiments were conducted on the Cozmo platform, which limits the robot’s action space and expressive range; future work should validate these findings on more capable robotic systems.

7 Conclusions & Future Work

This paper examined whether gaze, a simple, non-verbal social signal, can help an embodied robot communicate failure to a human collaborator during a object-search task. Specifically, we asked whether this signal could help the robot succeed even when it lacks the reasoning needed to solve the problem on its own. We first compared two planners that differ in capability: a rule-based planner that reacts to raw detections, and a vision-language model planner that reasons over full scene semantics. The VLM planner found the target object far more often (87.2%) than the rule-based planner (65.9%), confirming that semantic reasoning helps when a search grows difficult.

This capability gap set up the contrast at the

center of our design. Rather than treating the rule-based planner as simply a weaker version of the VLM planner, we treat it as a stand-in for an agent that has not yet developed full semantic and linguistic reasoning, much like a young child who cannot yet utter request for help due to a lack of vocabulary, but who still succeeds by drawing a caregiver into helping through emotional displays and eye gaze. This raised our central question: if the robot lacks that reasoning capability, can interaction with a human partner make up the difference instead? To test this, we added a gaze-based *stuck* signal to the rule-based planner and compared it against a verbal ‘give-up’ as the baseline, while holding mid-search GRED frustration displays constant across both conditions. We found that the rule-based planner reached 87.87% with gaze, matching the VLM planner outright, while the verbal baseline raised it only to 68.75%.

Gaze still worked even though participants did not always say they felt invited to help, which suggests it acted as a subtle cue rather than a direct request (Kompatsiari et al., 2018). Taken together, our results show that a robot does not need full linguistic or semantic capability to collaborate successfully with a human: interaction itself, prompted by a well-timed nonverbal signal, can recruit the help needed to make up for capability the robot does not have, in the same way that young children rely on gaze rather than language to get help from a caregiver (Plane et al., 2018).

For future work, we plan to test whether gaze will work with a VLM-planner that is trained to be more on par with child-level exploration in more visually cluttered environments, with a larger participant pool, and to compare gaze against an explicit verbal invitation baseline (e.g., “can you help me?”) rather than only a failure announcement such as “can’t find.” Importantly, the interaction should result in the robot not just finding objects, but exploring them such that it can also learn grounded representations of the objects connected to referring expressions made by humans to them, all learned through interaction.

Acknowledgements. We are grateful to the anonymous reviewers for their helpful feedback. This material is based upon work supported by the National Science Foundation under Grant No. 2429968 and 2140642. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not nec-

essarily reflect the views of the National Science Foundation.

References

- Henny Admoni and Brian Scassellati. 2017. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuezhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zhen Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-vl technical report](#). *Preprint*, arXiv:2502.13923.
- Rista Baral, Bethany Grenz, and Casey Kennington. 2025. Recognizing and generating novel emotional behaviors on two robotic platforms. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 21503–21510. IEEE.
- Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphization of robotic agents. *International journal of social robotics*, 1(1):71–81.
- Elizabeth Cha and Maja Matarić. 2016. Using non-verbal signals to request help during human-robot collaboration. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5070–5076. IEEE.
- Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, and 1 others. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.
- Kyveli Kompatsiari, Francesca Ciardo, Vadim Tikhonoff, Giorgio Metta, and Agnieszka Wykowska. 2018. On the role of eye contact in gaze cueing. *Scientific reports*, 8(1):17842.
- Matteo Lavit Nicora, Pooja Prajod, Marta Mondellini, Giovanni Tauro, Rocco Vertechy, Elisabeth André, and Matteo Malosio. 2024. Gaze detection as a social cue to initiate natural human-robot collaboration in an assembly task. *Frontiers in Robotics and AI*, 11:1394379.
- Anna Manaseryan, Porter Rigby, Brooke Matthews, Catherine Henry, Josue Torres-Fonseca, Ryan Whetten, Enoch Levandovsky, and Casey Kennington. 2025. rrSDS 2.0: Incremental, modular, distributed, multimodal spoken dialogue with robotic platforms. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Avignon, France. Association for Computational Linguistics.

Alexandra Weidemann and Nele Rußwinkel. 2021. The role of frustration in human–robot interaction—what is needed for a successful collaboration? *Frontiers in psychology*, 12:640186.

C Post-Block Questionnaire

Participants completed the following questionnaire after each block. All rating items use a 5-point scale. Endpoint labels are shown in parentheses.

Part 1: Overall experience

1. How many objects did the robot find on its own? (numeric)
2. How many objects did the robot find with your help? (numeric)
3. How interesting was the robot to interact with? (1 = Not at all; 5 = Very)
4. Would you like to spend more time with the robot? (1 = Not at all; 5 = Very much)
5. Cozmo found the object I asked it to find. (1 = Strongly disagree; 5 = Strongly agree)
6. I felt invited to participate when the robot seemed unsure what to do. (1 = Strongly disagree; 5 = Strongly agree)
7. I felt engaged with the robot, not just watching it. (1 = Strongly disagree; 5 = Strongly agree)
8. The robot's movements made sense for finding the object. (1 = Strongly disagree; 5 = Strongly agree)
9. The robot's movements made sense for when the robot was unable to find an object. (1 = Strongly disagree; 5 = Strongly agree)
10. I felt we were working together to achieve a shared goal. (1 = Strongly disagree; 5 = Strongly agree)

Part 2: Impression of the robot — Godspeed questionnaire (Bartneck et al., 2009) *Anthropomorphism*

1. How aware did the robot seem? (1 = Unconscious; 5 = Conscious)
2. How natural did the robot seem? (1 = Artificial; 5 = Lifelike)
3. How would you describe the robot's movement? (1 = Moving rigidly; 5 = Moving elegantly)

Animacy

1. How alive did the robot appear? (1 = Dead; 5 = Alive)
2. How lively did the robot appear? (1 = Stagnant; 5 = Lively)
3. How interactive was the robot? (1 = Inert; 5 = Interactive)

Perceived Intelligence

1. How competent did the robot seem? (1 = Incompetent; 5 = Competent)
2. How knowledgeable did the robot seem? (1 = Ignorant; 5 = Knowledgeable)
3. How intelligent did the robot seem? (1 = Unintelligent; 5 = Intelligent)
4. How sensible were the robot's actions? (1 = Foolish; 5 = Sensible)

Likability

1. Please rate your impression of the robot in this session. (1 = Unlikable; 5 = Likable)
2. Please rate your impression of the robot in this session. (1 = Unpleasant; 5 = Pleasant)

C.1 Part 3: Godspeed results

Godspeed subscales We also administered the Godspeed questionnaire (Bartneck et al., 2009), which measures four dimensions of robot perception on 5-point bipolar scales. *Anthropomorphism* (3 items: Unconscious–Conscious, Artificial–Lifelike, Moving rigidly–Moving elegantly) captures how human-like the robot feels. *Animacy* (3 items: Dead–Alive, Stagnant–Lively, Inert–Interactive) captures how alive and active it appears. *Perceived Intelligence* (4 items: Incompetent–Competent, Ignorant–Knowledgeable, Unintelligent–Intelligent, Foolish–Sensible) captures how capable it seems. *Likability* (2 items: Unlikable–Likable, Unpleasant–Pleasant) captures how pleasant it is to interact with. Each subscale score is the mean of its items; if an item is missing for a participant, it is excluded from that participant's mean.

No subscale differed significantly (Table 4). *Anthropomorphism*: Baseline = 2.42 (SD = 1.17), Gaze = 2.14 (SD = 0.74), $W = 10.5$, $p = 1.0$ ($N = 7$). *Animacy*: Baseline = 3.25 (SD = 0.99), Gaze = 2.96 (SD = 0.84), $W = 6.0$, $p = 0.813$,

$r = 0.181$. Perceived Intelligence: Baseline = 3.16 (SD = 0.63), Gaze = 3.19 (SD = 0.70), $W = 14.0$, $p = 1.0$. Likability: Baseline = 4.31 (SD = 0.75), Gaze = 4.63 (SD = 0.52), $W = 1.5$, $p = 0.375$, $r = 0.639$.

This is expected. Godspeed reflects participants’ overall impression of the robot across the full session, not just the stuck moment, so large between-condition differences are unlikely. Perceived Intelligence was nearly identical (3.16 vs. 3.19), showing that the gaze turn did not make Cozmo seem less capable. Anthropomorphism was low in both conditions, consistent with Cozmo being a small toy robot. Likability trended higher for Gaze ($r = 0.639$), the largest effect across all subscales, but did not reach significance at $N = 8$.

Subscale	Baseline		Gaze		W	p
	M	SD	M	SD		
Anthropomorphism [†]	2.42	1.17	2.14	0.74	10.5	1.000
Animacy	3.25	0.99	2.96	0.84	6.0	0.813
Perceived Intelligence	3.16	0.63	3.19	0.70	14.0	1.000
Likability	4.31	0.75	4.63	0.52	1.5	0.375

Table 4: Godspeed subscale means (1–5) and exact Wilcoxon statistics ($N = 8$ unless noted). [†] $N = 7$ for Gaze (P4 missing). Likability showed the largest effect ($r = 0.639$), favouring Gaze. No subscale reached significance ($\alpha = 0.05$).

D Figures

Figure 9 summarizes the ratings of participant on Task performance and Perceived Intelligence of the Base system between the VLM planner and Yolo planner (baseline).

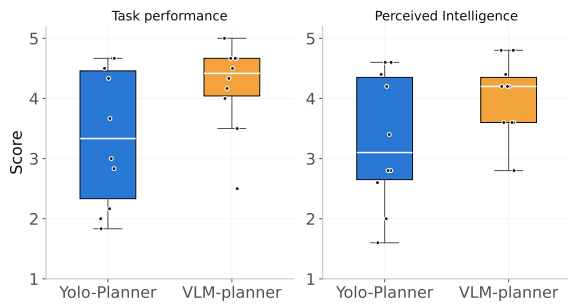


Figure 9: Task performance and Perceived Intelligence subscale by planner (Base system)