

While You're Still Busy: Context-Aware Approach Optimization for Proactive Mobile Robots

Ziru Wei
ziruw@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Di Wen
diwen@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Leyuan Chen
leyuanc@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Hassan Shahzad
zcabhs0@ucl.ac.uk
Robotics AI, UCL
London, United Kingdom

Alexandra Ion
alexandraion@cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

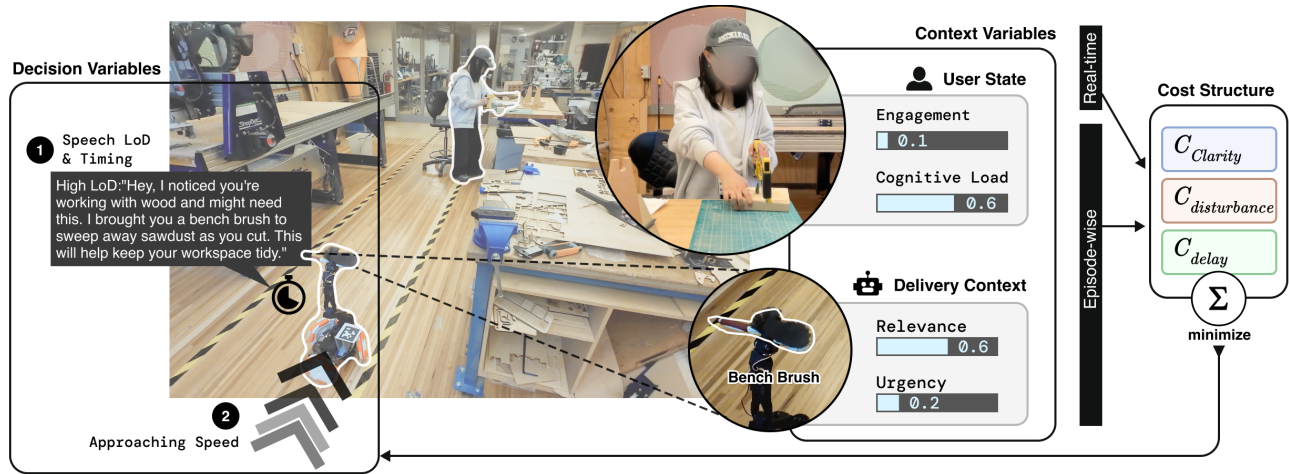


Figure 1: We present a system that considers user and delivery context to adapt the approaching behavior of mobile robots in real-time. The context variables include user state (engagement level, cognitive load) and object delivery context (relevance, urgency), which feed into a cost function that balances clarity, disturbance, and delay costs. The optimizer outputs two decision variables: speech timing & level of detail (LoD), and approaching speed. Here, low urgency and moderate cognitive load lead to a slower approach with a detailed explanation, together shaping a hesitant robot approach.

ABSTRACT

Robot assistants are increasingly capable of assisting users in everyday environments, yet they don't take users' cognitive context into consideration. As a result, proactive behaviors may interrupt users during cognitively demanding tasks, before they can acknowledge or respond. We identify this pre-acknowledgment window as a critical phase and introduce our real-time optimization approach that adapts how robots approach and communicate during this window, using user engagement signals and task context. Our system jointly optimizes movement and speech by solving structured, closed-form quadratic subproblems that balance disturbance, clarity, and efficiency online. This enables continuous adjustment of approach movement, timing of intent expression, and level of

detail without explicit user input. In a within-subjects ablation user study ($N = 12$), we find that adapting speech to context improves intent clarity, that speed adaptation alone has limited impact without paired speech adaptation, and that coordinating both channels yields the strongest animacy. We also demonstrate the real-world application in three everyday scenarios.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**.

KEYWORDS

Embodied Interaction, Physical AI, Adaptive Robot Behavior

ACM Reference Format:

Ziru Wei, Di Wen, Leyuan Chen, Hassan Shahzad, and Alexandra Ion. 2026. While You're Still Busy: Context-Aware Approach Optimization for Proactive Mobile Robots. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26)*, November 02–05, 2026, Detroit, MI, USA. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3830398.3830714>



This work is licensed under a Creative Commons Attribution 4.0 International License. *UIST '26, November 02–05, 2026, Detroit, MI, USA*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2856-3/2026/11.
<https://doi.org/10.1145/3830398.3830714>

1 INTRODUCTION

With AI revolutionizing our digital work, the next frontier is physical AI entering our physical spaces [41, 43, 52]. Beyond executing tasks, proactive robots can assist us in anticipating our needs that may slip our minds (e.g., forgetting keys when leaving the house, absentmindedly placing your hand next to dangerous objects, etc.) [23, 35].

To fully leverage these benefits in everyday physical assistance, we must carefully consider the robot’s interaction with the user in their environment. Compared to more common context-aware digital agents, embodied proactive robots carry higher interactional stakes and a greater risk of becoming obtrusive due to two main properties: First, **physicality**: unlike digital alerts, robot actions occupy space, require time to execute, and present as motion in the user’s environment, which can disrupt users’ spatial comfort and focus. Second, **intent**: intent is often explicit in screen-based interfaces but less transparent in human-robot interaction, so users must actively attend to the robot to understand its goal.

Consider the example of you being highly focused on studying for an exam. Your robot assistant enters the room, but you don’t have spare bandwidth to engage with it or try to understand what it intends to do. Should the robot proceed and approach you? Should it hesitate and wait for you to engage with it? What should it do if it has an urgent object to deliver to you, like your ringing phone receiving an interview callback? How should it proceed if it has a relevant book that may help with your study task? In this moment, an improper approach style, and ill-timed or poorly detailed intent clarification can easily turn a potentially supportive intervention into a distraction.

Prior work on proactive digital agents (often in extended reality setups) investigated interface adaptation by taking timing, modality, and context into consideration [36, 40, 54]. They generated interesting design implications, but they *can’t be readily applied* to physical assistants. On the other hand, prior systems for robot intervention has focused largely on *when* to interact [5, 10, 14]. The important question of *how* a robot should initiate the interaction based on context remains open.

In this paper, we address this gap by introducing context-aware optimization into physical human-robot interaction, presenting a practical design and system contribution through a real-time, deployable framework that optimizes a robot’s approach style.

1.1 PROACH: Proactive Approach Optimization

We introduce our system *PROACH*, which considers context and user state to adapt the approaching behavior of mobile robots in real-time. We structure the problem into *context* and *action* space, and map them through an optimization-based decision layer onto our robot behavior decision variables, as we show in Fig. 1.

In the *context* space, we include task-based variables, i.e., urgency, and relevance of the robot’s delivery to users’ current task, along with estimation of their cognitive load. Additionally, we continuously update implicit engagement signals from the user. As for the *action* space, we define the output as the robot’s moving speed and intent clarification (i.e., trigger speech output, level of detail of explanation) at any moment during the robot’s approach. Note that instead of computing a speed at the beginning of the

approach based on the user’s context then, we continuously update and compute appropriate speed and speech adaptation in real-time. Therefore, *PROACH* is a middleware optimizer that continuously adapts approach speed and speech detail as proximity and user attention evolve by solving the multi-objective optimization online.

The example in Fig. 1 shows how the user is working in their wood shop as the robot approaches them to bring a bench brush. The brush is relevant to the task, but not urgent since they may clean up later. Given that the task creates medium cognitive load and the user is not engaging with the robot (inferred from the upstream perceiving modules), the robot proceeds slowly to not disturb the user. As the user engages with the robot by looking in its direction, the robot explains its intent and adjusts its speed to bring the object faster after the user’s acknowledgement. We evaluate our robot approach optimization by performing a user study where we systematically vary the urgency and relevance of delivered items. Paired with our real-time engagement assessment using live video feedback, we find that adaptive behavior is perceived as more appropriate, clearer in intent, and more animate. In particular, jointly adapting speech content and movement yields the highest intent clarity, and participants attribute social qualities such as attentiveness and liveliness to the coordinated behavior. The novelty of our optimization framework lies in modeling task-dependent context, specifically the urgency and relevance of the delivered item, as well as jointly coordinating speech and movement outputs.

In summary, we make the following **contributions**:

- (1) We present an *optimization-based system* that adapts robot approach speed and speech style in real time from task context and engagement signals through multi-objective optimization.
- (2) We *evaluate* our system components in a user study and compare to three ablated versions. Our findings show that adaptive speech significantly improves intent clarity, while movement adaptation contributes to perceived animacy and appropriateness but does not reach statistical significance on its own.
- (3) We *demonstrate* our system both in a user study and through application examples spanning various domains, including offices, workshops, living rooms, and study halls.

2 RELATED WORK

Our work builds on prior work on proactive assistance and research on approach behaviors of robots. We also draw from the rich literature on computational methods for minimizing user disruption.

2.1 Proactive Assistance: What, When, and How

Proactive agents that anticipate user needs without explicit commands trace back to Horvitz’s mixed-initiative interaction [27] and attention-sensitive alerting [28, 29], and have accelerated with modern methods including large language model (LLM) and vision-language model (VLM) driven autonomy in open-ended settings [51]. Recent work on proactive agents has converged on three design axes: *what* to help with, via predictive user modeling [39] and proactive conversational support [42]; *when* to help, via working-memory modeling [54] and BDI-based inference for AR task guidance [38]; and *how* to deliver help, via content-modality reasoning adapted to

attentional load [36], LLM-driven everyday objects that reposition themselves [23], and smartwatch-based just-in-time procedural guidance [1]. These advances, however, are concentrated in digital and screen-based agents. Physical robot interaction remains largely command-driven, and where robots do act proactively, the higher stakes of embodiment make the risk of obtrusiveness far greater [61]. The digital proactive paradigm offers valuable insights on what, when, and which modality to act in [36, 38, 54], and physical proactive assistance has begun to emerge [10, 23, 72], but coordinating movement and communication for a physical robot introduces challenges that these frameworks do not address.

2.2 Robot Approach Behavior

Hall's proxemic zones [22] have been extensively mapped to HRI, establishing how factors such as robot speed, approach direction, gaze, and individual traits shape preferred approaching endpoints [16, 49, 56, 64, 69]. Based on proxemics theory, methods such as the Social Force Model [18, 25] and interaction potential fields [47] have been widely adopted for socially aware navigation, while approach-specific systems optimize paths based on real-time human state to respect personal space and reduce obtrusiveness [34, 57]. Yet these contributions characterize *where* a robot should decelerate and end up, and *which path* it should take, not how it should coordinate behavior *during* the approach.

Prior work establishes that effective robot approach should unfold temporally across multiple channels. Speed and proximity significantly shape how users interpret robot intent: Saulnier et al. showed that even minimal nonverbal variations during approach are read as social signals [58], and Rousseau et al. found that multi-modal behavior is preferred when robots initiate interaction [55]. Beyond speech, other nonverbal channels such as robot gaze, indicator lights, and motion velocity have also been incorporated as critical outputs during approach [26]. Building on this, Shi's group developed approach systems that coordinate locomotion, arm extension, and verbal cues within a single episode [33, 59], though speech or gesture remains secondary to locomotion in their pipeline. Besides, Mead and Matarić showed that approach trajectories themselves can be designed to facilitate multi-modal interaction opportunities [47].

These works establish that multi-modal approach behavior matters, but nuanced joint control of movement and speech and task-aware adaptation remain unexplored. We build on this foundation by jointly adapting movement and speech in real time and grounding the adaptation in the user's ongoing task context.

2.3 Computational Methods for Minimizing User Disruption

Minimizing disruption to users has a long history in HCI, where computational methods have been developed to manage the cognitive cost of system-initiated actions. Decision-theoretic models first optimized *whether* to interrupt [28–30], and task-structural breakpoints refined *when* [31]. Later systems moved beyond timing into *presentation*: SUPPLE formulated UI layout as constrained effort minimization to reduce cognitive load [19, 20], Lindlbauer et al. adapted Mixed-Reality detail levels to cognitive load in real time [40], and Cheng et al. incorporated virtual-physical semantic

associations [13]. Most recently, ProMemAssist modeled intervention value against cognitive cost for proactive assistance timing [54]. These methods provide valuable insights, but are difficult to transfer directly to physical agents, where disruption is not a discrete event but a continuous process that unfolds in a shared space.

Within HRI, where this transfer is most needed, work on minimizing disruption has likewise focused on a binary "go-or-wait" decision. Early methods used spatial engagement classifiers [9, 53], followed by temporal interruptibility models [5, 17], RL-based personalization of intervention timing [14], and more recently VLM-based context assessment for deciding when to initiate interaction [70, 74]. Decision-theoretic reasoning has also been applied to communication timing in human-robot collaboration [66]. Mannem et al. empirically quantified how proximity, timing, and content shape perceived disruption cost in physical robot interruptions [45], demonstrating that these dimensions are rich enough to warrant optimization. Our work takes a step forward: treating disruption as a *continuous spatial process* that evolves with proximity and is managed in real time through coupled movement and speech, during the pre-acknowledgment window in which no explicit user input is available.

3 SYSTEM IMPLEMENTATION

Our implementation is guided by three **design goals**. While each goal has been recognized individually in interruption management, proactive agent design, and human-robot interaction [46, 54, 69], they are coupled and cannot be satisfied independently in the context of robot approach: conveying intent conflicts with staying unobtrusive, explaining slows delivery, and politeness conflicts with being efficient. Pilot observations confirmed each trade-off (Fig. 2), motivating the structure of our joint optimization formulation.

- G1: **Minimize Disturbance.** When the user's cognitive load is high, robot behaviors should become more conservative. For example, fast movement and lengthy speech should be avoided, especially as the robot gets closer.
- G2: **Provide Timely Assistance.** When the delivery is urgent, the robot should prioritize progress toward completing the handoff promptly.
- G3: **Maintain Intent Legibility.** The robot should make its intent understandable. More explicit explanations are needed when the assistance is more likely to interrupt or redirect the user's ongoing activity.

We thus formulate behavior selection as an optimization problem, allowing the robot to jointly regulate movement assertiveness and communicative detail under the context.

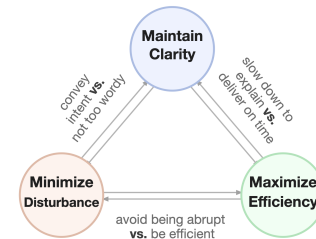


Figure 2: Trade-offs between the three design goals.

Table 1: Input variables as the context space of the system.

Input type	Var.	Description	Range
User State	x_{load}	Cognitive load of the user's current activity	Continuous in [0, 1]
	x_{engage}	Real-time user engagement and attention level	Continuous in [0, 1]
Delivery Context	$x_{urgency}$	Urgency of the robot's delivery task	Continuous in [0, 1]
	x_{relev}	Relevance of delivery to the user's current activity	Continuous in [0, 1]

3.1 Inputs and Outputs

3.1.1 Inputs as the Context Space. PROACH takes four contextual inputs along two dimensions (see Table 1). *User state*, comprising cognitive load from the primary tasks and engagement level with the robot, captures how receptive the user currently is. *Delivery context*, comprising urgency and relevance, characterizes what the robot is bringing and why it matters.

User State. Cognitive load and engagement jointly describe the user's current capacity and willingness to receive an interruption. Cognitive load, drawn from interruption management research [46, 63], reflects the demand of the ongoing activity; reduced processing capacity amplifies sensitivity to any interruption [3, 63]. In our implementation, cognitive load is a task-specific factor that reflects the overall workload. Specifically, we use a predefined value for cognitive load in the user study (Section 4) to keep conditions comparable. To demonstrate the entire system, we use LLM-inferred values in Section 5, since recent work has demonstrated that LLMs can successfully estimate cognitive load in real-time [71].

Engagement level characterizes the degree to which a user maintains a perceived connection with a robot [62], and has been modeled in HRI through spatial proximity [48] and behavioral cues such as gaze and head orientation [5, 17]. We adopt head-pose-derived engagement as the real-time complement to task-level cognitive load, capturing dynamic shifts in user attention. Using an environment-mounted camera in front of the user, we estimate engagement online by mapping yaw and pitch from a 3D face alignment model [21]¹ to a continuous score in [0, 1], where looking away yields lower engagement and turning toward the robot increases it. The signal is smoothed with an exponential moving average and thresholded at 0.4 to derive a discrete engaged state for event counting. This provided a stable signal in our current system and can be replaced by more advanced approaches in future work.

Delivery Context. Urgency and relevance characterize the robot's delivery relative to the user's current activity. Interruption management research identifies both the important alert and the task relevance to the user's ongoing task as key factors shaping how interruptions should be presented [46, 50]; we adopt these as $x_{urgency}$ and x_{relev} . Again, we use predefined values for the controlled study in Section 4, and an LLM-inferred value for the demonstration in Section 5

Table 2: Output variables and behavior mapping.

Var.	Description	Range	Behavior Mapping
s	Approach speed	[0,1]	Mapped to robot velocity; values below 0.05 are set to zero.
l	Explanation LoD	[0,1]	[0, 0.40): keep silent [0.40, 0.50) - low LoD: item + urgency, ~2 s [0.50, 0.75) - mid LoD: low LoD + actionable advice, ~4 s [0.75, 1] - high LoD: mid LoD + consequence, ~9 s

3.1.2 Outputs as the Action Space. We scope the action space to two channels that the robot can modulate continuously during approach: movement and speech. Approach speed governs both physical assertiveness in shared space [69] and nonverbal intent signaling [65, 73], while explanation level of detail (LoD) governs how much attentional capacity the robot claims through speech [4, 12, 40]. At each decision step, the optimizer solves for normalized speed $s \in [0, 1]$ and speech LoD $l \in [0, 1]$, mapped to executable behaviors as summarized in Table 2. In our implementation, normalized speed is converted to robot velocity (max = 0.44 m/s), and instantiations across LoDs are included in the Supplementary Material.

3.2 Formulation

Following prior optimization-based approaches to context-aware adaptation [19, 40, 54], we formulate behavior selection as a light-weight optimization that balances competing objectives under the current context. The formulation has four components: (1) hard constraints that govern speech admissibility; (2) a weighted cost structure balancing clarity, disturbance, and delay; and (3) regularization and smoothing for coherence and physical comfort. We refined the weights based on empirical evidence from internal lab members ($N = 5$) during the system design, following prior work [13, 40, 54].

The decision process is composed of two coupled box-constrained subproblems; in practice, each subproblem is solved over [0,1]:

$$l^* = \arg \min_{l \in [0,1]} f_l(l), \quad s^* = \arg \min_{s \in [0,1]} f_s(s; l_{used}).$$

To rule out overlapping utterances and repeated re-explanations, we impose two hard constraints. First, a *Speech Exclusivity Constraint* prevents the robot from initiating a new utterance while a previous utterance is still ongoing. Second, a *Single-Opportunity Speech Constraint* allows at most one utterance during each delivery; once intent has been communicated, the LoD variable is locked to zero for the remainder of the approach.

3.2.1 Cost Structure. We optimize a weighted sum of three terms:

$$\min_{s,l} w_c C_{clarity}(s, l) + w_d C_{disturbance}(s, l) + w_t C_{delay}(s, l),$$

where $s \in [0, 1]$ denotes normalized speed and $l \in [0, 1]$ denotes explanation level of detail. Fig. 3 summarizes the role of each term on the two action variables.

¹We use the TensorFlow Lite reimplementation: <https://github.com/1996scarlet/Dense-Head-Pose-Estimation>.

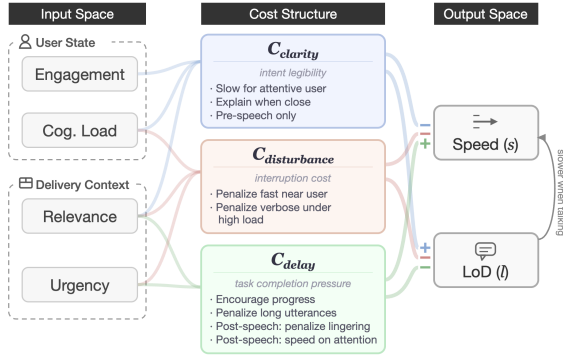


Figure 3: Role of each cost term on the two action variables.

In the implementation, LoD is solved first and then coupled into the speed objective through a one-way walk-talk penalty:

$$l^* = \arg \min_{l \in [0,1]} f_l(l), \quad s^* = \arg \min_{s \in [0,1]} (f_s(s) + \gamma l_{\text{used}}^2 s^2),$$

where l_{used} is the effective speech level used for movement suppression. This design allows speaking to slow the robot down, while speed does not directly suppress the tendency to explain. Around and during speech, a bi-directional speed adjustment further modulates speed based on engagement direction: if the user is looking at the robot, it slows down; if the user looks away, it speeds up. This adjustment ramps in as the LoD solution rises before speech is triggered, remains active during speech, and decays gradually afterward.

Clarity term. The clarity term encourages the robot to make its intent understandable. We model it as a linear preference over both motion and explanation:

$$C_{\text{clarity}}(s, l) = \beta_s^{(c)} s + \beta_l^{(c)} l.$$

The speed coefficient $\beta_s^{(c)} = w_c \cdot x_{\text{engage}}$ penalizes forward motion in proportion to user engagement, discouraging the robot from rushing toward an attentive user before explaining itself. The explanation coefficient combines a proximity-driven and an engagement-driven component:

$$\beta_l^{(c)} = \underbrace{-w_c \phi(d_{hr})}_{\text{proximity push}} + \underbrace{-w_c x_{\text{engage}} \rho(x_{\text{revel}})}_{\text{engagement push}},$$

where $\phi(d_{hr}) = \text{clamp}\left(\frac{d_{\text{explain}} - d_{hr}}{d_{\text{explain}} - d_{\text{critical}}}, 0, 1\right)$ ramps linearly as the robot enters conversational range, and $\rho(x_{\text{revel}}) = 1 + k_p (1 - x_{\text{revel}})$ amplifies the push when task relevance is low. The proximity component drives the robot toward speaking as it nears the user, independently of engagement. The engagement component activates when the user attends to the robot, with strength scaled inversely by relevance: information congruent with the user's current task is easier to integrate [67], so a relevant delivery is more self-evident and demands less explanation. Additionally, a baseline clarity component pushes LoD upward for low-relevance deliveries even without engagement, ensuring the robot still explains externally motivated deliveries. Cognitive load suppresses these components, so that high-load users receive less verbose explanations.

Disturbance term. The disturbance term penalizes obtrusive behaviors in both movement and speech. Its base form is quadratic:

$$C_{\text{disturbance}}(s, l) = \alpha_s^{(d)} s^2 + \alpha_l^{(d)} l^2.$$

The movement coefficient is a multiplicative composition of context factors:

$$\alpha_s^{(d)} = w_d \cdot \hat{x}_{\text{load}} \cdot \psi(d_{hr}) \cdot \delta_s(x_{\text{revel}}) \cdot \frac{1}{1 + k_u x_{\text{urgency}}},$$

where $\hat{x}_{\text{load}} = k_{\text{base}} + (1 - k_{\text{base}}) x_{\text{load}}$ is cognitive load with a baseline floor ensuring a minimum disturbance sensitivity even for low-load tasks, $\psi(d_{hr})$ is a proximity amplifier that grows as the robot enters the proxemic zone, consistent with physiological evidence that robot motion speed elevates sympathetic arousal [24], and δ_s is a mild relevance discount. Urgency reduces but does not eliminate disturbance sensitivity, so even time-critical deliveries respect the user's spatial comfort.

The explanation coefficient $\alpha_l^{(d)} = w_d \cdot x_{\text{load}} \cdot \delta_l(x_{\text{revel}})$ penalizes detailed speech as more intrusive, scaling directly with cognitive load. $\gamma l^2 s^2$ further penalizes simultaneous movement and speech; l is solved first and injected as a constant into the speed quadratic, enabling speech to suppress speed. The coupling strength γ is discounted by urgency, so that time-sensitive deliveries tolerate more concurrent motion. In the post-speech stage, $\alpha_s^{(d)}$ is discounted because lingering itself becomes the more salient disturbance based on empirical evidence.

Delay term. The delay term penalizes slow task progress and unnecessarily long explanations:

$$C_{\text{delay}}(s, l) = \beta_s^{(t)} (1 - s) + \alpha_l^{(t)} l^2.$$

The speed coefficient combines effective urgency with distance and relevance scaling:

$$\beta_s^{(t)} = w_t \cdot \hat{x}_{\text{urgency}} \cdot \eta(d_{hr}) \cdot \sigma(x_{\text{revel}}),$$

where $\hat{x}_{\text{urgency}} = \text{clamp}(x_{\text{urgency}} + k_{\text{boost}} (1 - x_{\text{load}}), u_{\text{min}}, u_{\text{max}})$ is an effective urgency that is boosted under low cognitive load and floored to ensure a baseline forward drive even for non-urgent deliveries. $\eta(d_{hr})$ attenuates $\beta_s^{(t)}$ as the robot enters the proxemic zone, as there is less need to hurry perceptually. $\sigma(x_{\text{revel}})$ scales down the drive for less relevant deliveries: items with lower task relevance warrant less time pressure. After speech has been delivered, two additional terms activate: a quadratic completion cost $k_c (1 - s)^2$ penalizes lingering, and an engagement-driven term reverses sign, treating user attention as a signal that intent has been conveyed and the robot can accelerate toward completion. The explanation coefficient $\alpha_l^{(t)} = w_t \cdot \hat{x}_{\text{urgency}}$ penalizes speech duration in proportion to effective urgency, so that time-sensitive deliveries favor briefer explanations.

3.2.2 Regularization and Smoothing. To avoid unstable frame-to-frame fluctuations in LoD, we apply a quadratic inertia term to the LoD objective:

$$f_l(l) \leftarrow f_l(l) + \lambda_{\text{lod}} (l - l_{\text{prev}})^2,$$

smoothing the trigger timing while preserving the steady-state optimum. After speech, the walk-talk speed suppression is released via exponential decay rather than removed immediately. Two terms are added to the speed objective independently of the cost weights:

a load-scaled s^2 baseline resistance that biases toward conservative speeds, and a distance-dependent s^2 proximity brake informed by proxemic comfort zones [69]. The resulting speed command is passed through an acceleration-limited smoother ($\pm 0.2 \text{ m/s}^2$) before execution.

4 EVALUATION

To understand how each dimension of adaptive robot approach behavior contribute to user experience, we conducted a within-subjects ablation study. We progressively add speech and movement optimization capabilities to assess the contribution of each component.

4.1 Participants

The study included $N = 12$ adult participants (10 female, 2 male; age $M = 24.8$, $SD = 1.7$, range 21–28) recruited from the local community. Most had limited experience with robots: six had interacted a few times, three once or twice, two never, and one regularly. Participants gave informed consent and received \$15 in cash or a gift card.

4.2 Tasks

The study followed an interruption paradigm commonly used in HCI and HRI [5, 46], in which participants performed a continuous primary task while robot deliveries acted as repeated interruptions.

4.2.1 Users' Primary Task. We used card sorting as the primary task because it occupies the user both physically and cognitively, and has been used in interruption research [2, 8]. We sampled 24 cards per session from 60 established items [8] with minimal repetition, creating four constant-difficulty variants. The task consisted of a sorting phase (10 min) followed by an organizing phase (2 min). During sorting, participants grouped printed cards and performed physical operations such as hole-punching and highlighting, which created demand for robot-delivered tools. During organizing, participants labeled and sealed the sorted groups. This two-phase task structure supported the urgency manipulation: some robot-delivered items were immediately useful during sorting, whereas others were only needed later during organizing.

4.2.2 Robot Delivery Interventions. We discretized urgency and relevance into high and low levels to cover the context space; the study evaluates the optimizer's formulation, not the full continuous range. Deliveries crossed urgency and relevance (see Fig. 4): high-relevance items supported the primary task (e.g., a puncher, binder rings), low-relevance items served external goals (e.g., a package to sign, water). High-urgency items required immediate attention; low-urgency items could be deferred. Each delivery drew from a pool of alternative items with the same attribute to avoid repetition.

4.3 Conditions

We developed four system variants that participants tested, one system per session. The systems are illustrated in Fig. 5.

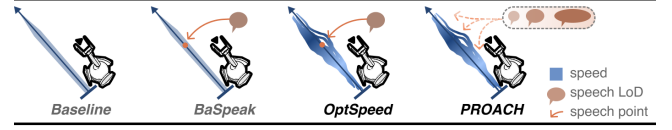


Figure 5: The four system variants in our ablation study.

The closest *Baseline* system we found uses no speech and a fixed speed profile based on the Social Force Model (SFM) [18, 25], in which speed decreases monotonically with distance to the user. This represents a common approach strategy in prior HRI work. Since our PROACH system modulates speed, speech trigger, and level of details, we add ablated system versions to understand their impact. Our *BaSpeak* variant uses the same speed modulation as *Baseline*, but adds a fixed speech component: a minimal utterance triggered at $d = 1.5 \text{ m}$, following distance-based initiation conventions [22, 32, 57, 60]. This isolates the contribution of speech presence. Our *OptSpeed* variant uses the same fixed speech as *BaSpeak* but replaces the SFM with our optimizer's adaptive speed. This isolates the contribution of adaptive movement. The full *PROACH* system jointly optimizes both speech and speed in real time based on context, as described in Section 3.2. To maintain comparability across conditions, we used predefined values for urgency, relevance, and cognitive load (see Fig. 4). User engagement is modeled dynamically to validate the system's real-time adaptation capability.

4.4 Procedure

After a background survey and familiarization trial, participants completed four sessions, one per system. As illustrated in Fig. 4, each session consisted of four deliveries from four distinct categories presented in a randomized order. Immediately after each delivery, participants completed a brief post-delivery survey before resuming the primary task. Each of the four sessions concluded with a post-session survey and a short semi-structured interview. Upon completion of all sessions, an exit interview gathered overall feedback across all conditions. To mitigate order effects, we counterbalanced all conditions, delivery object sets, and session orders across participants, and randomized the delivery order before beginning the study.

4.5 Study Setup and Apparatus

The study took place in a $1.6 \times 3.6 \text{ m}$ lab space configured as a personal workspace, with participants seated at a desk, as shown in Fig. 4. The robot approached from a distance of 3.2 m along a straight-line path on the participant's right-hand side and retreated behind a curtain after each delivery or rejection. The experimenter manually reloaded the next item out of view to preserve perceived robot autonomy. A simple spoken override channel is incorporated for direct user acceptance or rejection, driven by a set of predefined keywords.

The study used a LeKiwi mobile manipulator with a SO-101 arm [11], equipped with a Bluetooth speaker for speech output. The environment included a webcam for AprilTag-based localization, a user-facing webcam for engagement detection, and a wireless microphone for spoken user commands. A robot-mounted Raspberry Pi 5 handled low-level control, while a local MacBook

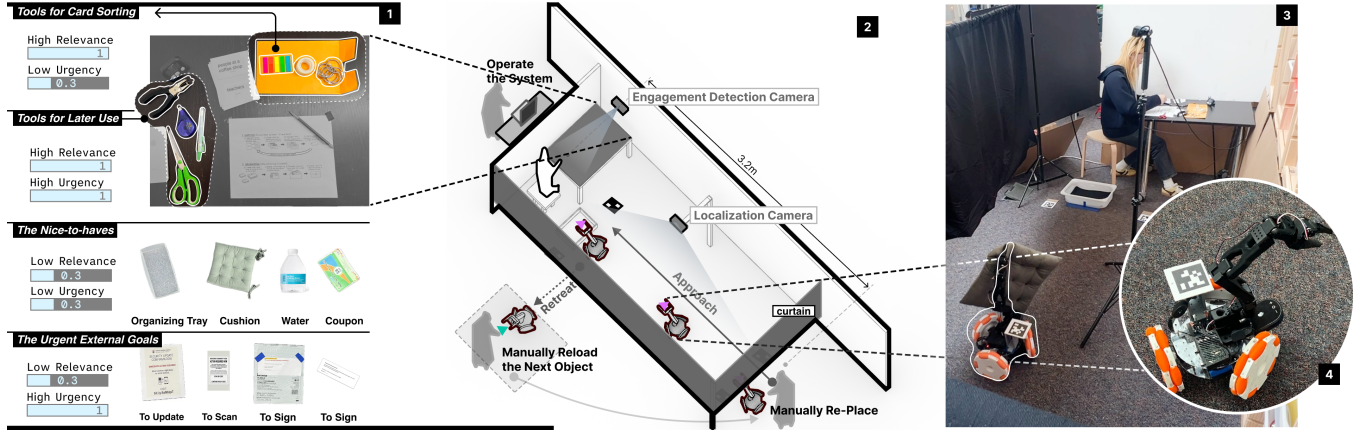


Figure 4: Overview of the study setup. (1) Delivery items organized by urgency and relevance; high-relevance items are placed on the participant’s desk (except the current delivery) to maintain task context. (2) The study environment layout. (3) A participant performing the primary task during a robot delivery. (4) The LeKiwi mobile manipulator.

Pro M1 ran perception and optimization. Engagement estimation and localization updated at approximately 30 Hz, the optimizer at approximately 7 Hz, and execution commands at 30 Hz. The system plays pre-synthesized ChatterboxTTS MP3 files for the robot’s audio output at runtime.

4.6 Data Collection and Analysis

Participants rated appropriateness, intent clarity, movement, and timing after each delivery (7-point Likert). Post-session, we collected Godspeed Animacy and Safety scales [6] and conducted interviews. Following insights from the HCI community [68], we analyzed session-level data ($N = 48$) using LMMs (lme4 [7]) and ordinal delivery-level data ($N = 192$) via CLMMs (ordinal [15]), with *system* as a fixed effect and *participant* as a random intercept. Significance was assessed via likelihood-ratio tests and Bonferroni-corrected comparisons (emmeans [37]).

4.7 Results and Findings

The main takeaways are that adaptive speech content acts as the primary driver of quantitative improvements in intent clarity and perceived appropriateness, while the standalone quantitative benefit of adaptive speed is less distinct. However, qualitative evidence strongly indicates that the unique value of movement lies in its joint coordination with speech, which prevents behavioral misalignment and establishes cohesive interaction. Since none of the participants expressed explicit commands to the robot, this validates our motivation that the ‘pre-acknowledgment window’ matters when interacting with a proactive robot. We detail our results in the following.

(1) Perceived animacy, safety, and appropriateness. All three proposed systems were perceived as substantially more animate than the silent baseline (see Fig. 6). *PROACH* was perceived as most animate ($M = 3.64$, $SD = 0.69$), followed by *OptSpeed* ($M = 3.25$, $SD = 0.88$) and *BaSpeak* ($M = 3.00$, $SD = 0.38$), while *Baseline* ($M = 1.97$, $SD = 0.70$) was markedly less animate. The animacy subscale

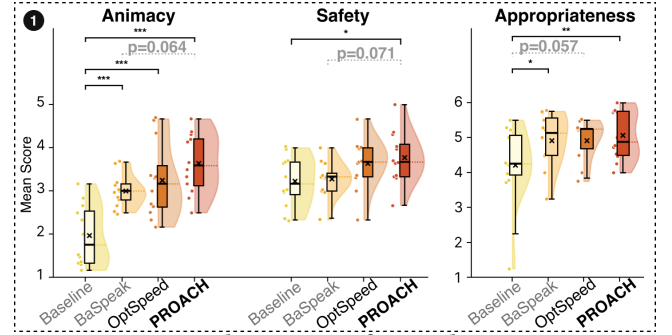


Figure 6: Perceived animacy, safety, and appropriateness.

showed high internal consistency ($\alpha = 0.915$). All three speech-enabled systems were perceived as significantly more animate than *Baseline* (*BaSpeak*: $t(33) = -4.36$, $p < .001$; *OptSpeed*: $t(33) = -5.42$, $p < .001$; *PROACH*: $t(33) = -7.06$, $p < .001$), and *PROACH* was marginally more animate than *BaSpeak* ($p = .06$), while no other pairwise differences reached significance.

Interview data suggest two contributing mechanisms. First, added speech elevated animacy. P04 noted longer utterances made the robot feel “more alive,” while P09 described the silent *Baseline* as “more apathetic.” Second, joint optimization produced a compound effect in *PROACH*: P08 noted that “the verbal explanation and the speed aligned very well,” producing an impression of the robot as “very lively,” the alignment of speech and movement, rather than either alone, yielded the strongest animacy signal. Perceived safety scores did not yield reliable differences; the low internal consistency of the subscale ($\alpha = 0.223$) precludes meaningful interpretation.

For perceived appropriateness, *PROACH* ($M = 5.06$, $SD = 1.04$; $z = -3.22$, $p = .008$) and *BaSpeak* ($M = 4.92$, $SD = 1.32$; $z = -3.02$, $p = .015$) were rated significantly more appropriate than *Baseline* ($M = 4.21$, $SD = 1.57$), while *OptSpeed* ($M = 4.92$, $SD = 1.01$; $z = -2.60$, $p = .057$) remained marginal. Notably, *BaSpeak* reached significance with fixed speech and fixed speed, while *OptSpeed* remained marginal despite adding adaptive speed with the same fixed

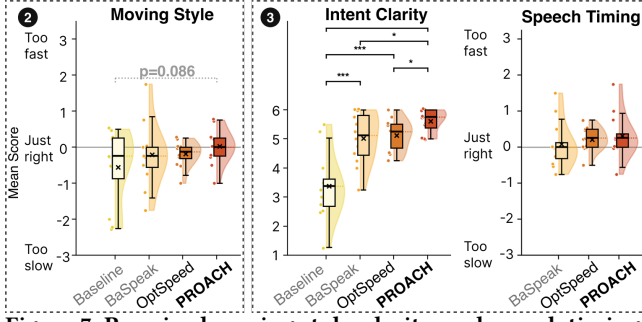


Figure 7: Perceived moving style, clarity, and speech timing.

speech. This suggests that appropriateness benefits from coherence between movement and speech: either both are consistent, or both adapt together. Interview data support this interpretation. Participants valued the behavioral consistency of *BaSpeak*: P09 noted that “consistency makes the robot’s behavior fall within expectation and felt calmer”, and P02 described constant behavior as “predictable.”

In contrast, *OptSpeed*’s mismatched adaptation was viewed as a negative aspect: P06 described the interaction as “in a rush” when the robot moved fast with a brief explanation for an urgent package delivery, suggesting that adapting speed without also adapting speech disrupts the coherence. Without joint optimization, a uniform movement speed profile paired with speech near the user was criticized for forcing users into “waiting for the robot” (P04, P07), with some noting it “held attention for longer and was especially annoying” (P05). For *PROACH*, users noted that the explanations “covered the whole approach distance,” eliminating the feeling of waiting (P04). Crucially, in the joint optimization, the dynamic movement provided a vital “buffer time to decide whether to keep the item or reject it” (P03). The joint optimization made the interaction feel “lively and efficient” (P07), “more humanoid” (P02), “intentionally helping” (P10), “less disruptive” (P03, P08), and “less intrusive” (P11). As reported in the animacy results, jointly adapting both channels builds on coherence, with participants describing the coordinated behavior as both clearer and more lively, which may contribute to the strongest effect of *PROACH*.

(2) Movement style. No significant effect of system was found for movement style ($p = .101$) (see Fig. 7). We found that *PROACH* was only marginally better rated than *Baseline* ($z = -2.45$, $p = .086$). The effect may have remained marginal partly because several participants did not reliably perceive speed variation (P10, P11), and partly because subjective speed perception tracked task context more than actual velocity, e.g., P02 and P09 perceived the robot as “too slow” when awaiting urgent items despite constant speed. This indicates that the robot’s maximum speed of 0.44 m/s is not sufficient for such tasks and future system designers should increase robot speed capabilities.

(3) Perceived intent clarity and speech timing. Speech substantially improved perceived intent clarity across all three speech-enabled systems (see Fig. 7). *PROACH* achieved the highest intent clarity ($M = 5.60$, $SD = 0.61$), followed by *OptSpeed* ($M = 5.13$, $SD = 0.98$) and *BaSpeak* ($M = 5.02$, $SD = 1.30$), with *Baseline* ($M = 3.38$, $SD = 1.62$) rated lowest. All three speech-enabled

systems were perceived as significantly clearer in intent than *Baseline* (*BaSpeak*: $z = -6.22$, $p < .001$; *OptSpeed*: $z = -6.08$, $p < .001$; *PROACH*: $z = -7.89$, $p < .001$). Among them, *PROACH* significantly outperformed both *BaSpeak* ($z = -2.72$, $p = .039$) and *OptSpeed* ($z = 2.94$, $p = .020$), whereas *BaSpeak* and *OptSpeed* did not differ ($z = 0.21$, $p = 1.0$).

Intent clarity benefits most from scaling speech detail to how self-evident the delivery is: relevant items need minimal explanation, while externally motivated ones require explicit justification. Participants independently articulated this principle: P11 stated the robot should “speak more for external goals” but “less or not at all for internal goals,” by which they meant task-relevant goals. Others described relevant items as self-explanatory while flagging irrelevant urgent deliveries as confusing or even rude without sufficient context (P01, P05, P06, P07).

Interestingly, we found no significant differences in vocal timing ($p = .155$). Individual preferences may matter more than contextual timing adaptation in this dimension, as some participants consistently preferred early speech (P01, P05, P09, P12) while others preferred it closer to arrival (P06, P07, P10). Additionally, the traveled distance in our study may have been too short to elicit differences and could be studied further.

4.7.1 Representative System Outputs. Fig. 8 illustrates the delivery timelines for one representative participant (P03). *Baseline* and *BaSpeak* produce fixed speed profiles regardless of engagement, while *OptSpeed* and *PROACH* modulate speed in response to real-time engagement signals. Speech behavior further differentiates the systems: *BaSpeak* and *OptSpeed* trigger speech at a fixed distance with a preset level of detail, whereas *PROACH* selects both timing and detail level adaptively. In **Stage 1 (Pre-speech, 0–6.75 s)**, high relevance but low urgency and no user attention lead the optimizer to settle at a steady cruising speed (~ 0.57), which proximity braking then gradually reduces to ~ 0.44 as the robot enters the proximity zone (~ 1.5 m). In **Stage 2 (During speech, 6.75–11.15 s)**, the user still shows zero engagement, but the proximity-driven clarity cost crosses the speech threshold at $d_{hr} = 1.06$ m, triggering a Mid-LoD utterance whose walk-talk penalty suppresses speed to ~ 0.29 ; mid-way through ($t \approx 8.9$ s) the user looks up (engagement $0 \rightarrow 0.93$), yet the optimizer responds with only a slight speech-speed boost as the walk-talk penalty remains dominant. In **Stage 3 (Post-speech, 11.15 s–end)**, the walk-talk penalty decays and a completion cost activates, together ramping speed from 0.29 back to ~ 0.59 ; when the user engages a second time ($t \approx 12.6$ s) the optimizer drives speed up to 0.77, but the final-approach deceleration layer ($d_{hr} < 0.20$ m) overrides it, forcing a linear ramp-down to zero at arrival.

5 APPLICATIONS

While task context was predefined in our controlled study, the real-world deployment requires automated inference. To this end, we extend *PROACH* using Gemini 2.5 Flash to infer user actions, tasks, and cognitive load from video stream at the task level. A separate LLM scores relevance and urgency $\in [0, 1]$, with candidate explanations generated via Google Cloud TTS with a latency of 200–500 ms. We evaluate it by demonstration across three scenarios (Figs. 9–11). In-depth evaluation of this upstream component remains future

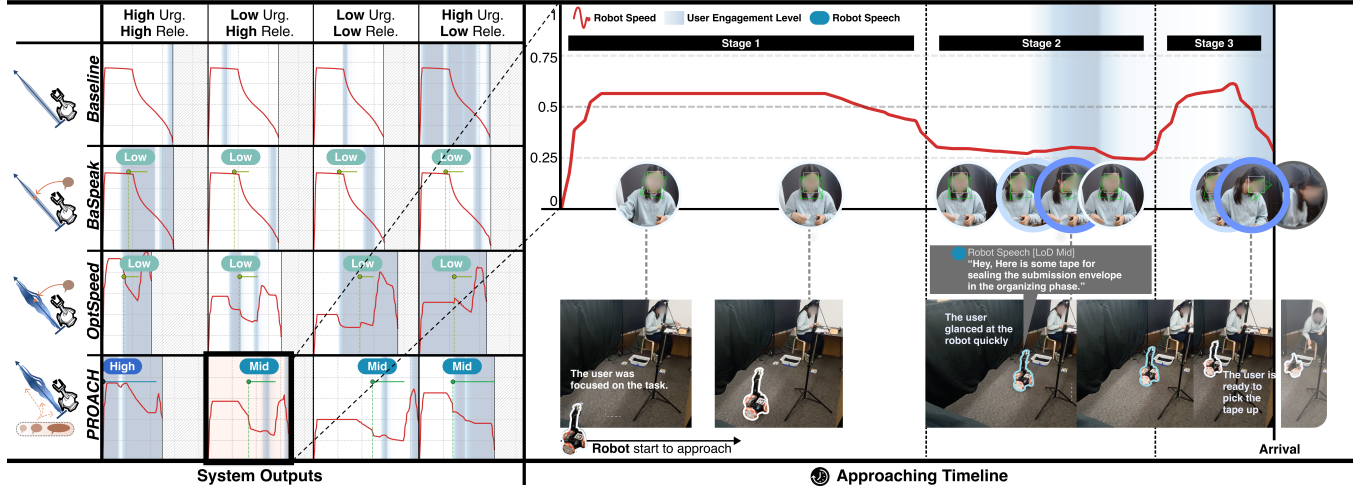


Figure 8: Visualization of robot approaching timelines for one participant. The red curve shows robot speed, blue shading shows real-time user engagement, and colored annotations show speech with level of detail. Left: overview across systems (rows) and delivery categories (columns); right: a detailed view of how the delivery unfolds within one example condition.

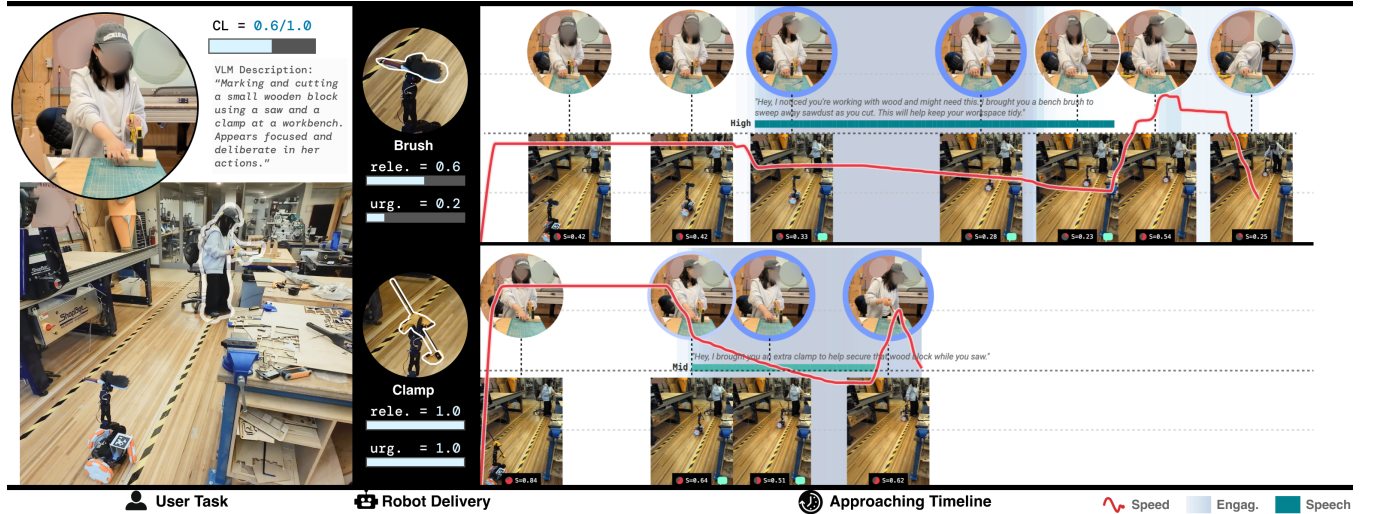


Figure 9: PROACH adapts delivery behavior to delivery context. Given the same woodworking task, the bench brush with lower relevance and urgency receives a slower approach with more detailed speech (top), while the clamp triggers a faster approach with briefer speech (bottom).

work, as it is not a core contribution of this paper. We describe how the system adapts to different aspects in the following sections.

5.1 Adapting to Object Context

The same user and task can produce different robot behavior depending on the object being delivered (Fig. 9).

Sarah is at her workbench, sawing and shaping a wood block for a landscape model, a task that tolerates interruption but does not invite it ($CL = 0.6$). When the robot arrives with a *bench brush*, useful for sweeping sawdust but not something Sarah is actively waiting for, it approaches at a moderate pace and explains its intent: “Hey, I noticed you’re working with wood and might need this. I brought you

a bench brush to keep your workspace clear.” After considering for a while, Sarah thinks that it’s nice to have, and takes the brush.

When instead the robot delivers a *clamp*, the situation changes, since the wood block needs securing before Sarah can safely continue cutting. The robot approaches noticeably faster and keeps its explanation more brief: “Hey, I brought you an extra clamp to help secure the wood while you saw.” Sarah reaches out and takes it right away.

5.2 Adapting to User State

The same task context can produce different robot behavior depending on the user’s momentary state (Fig. 10).

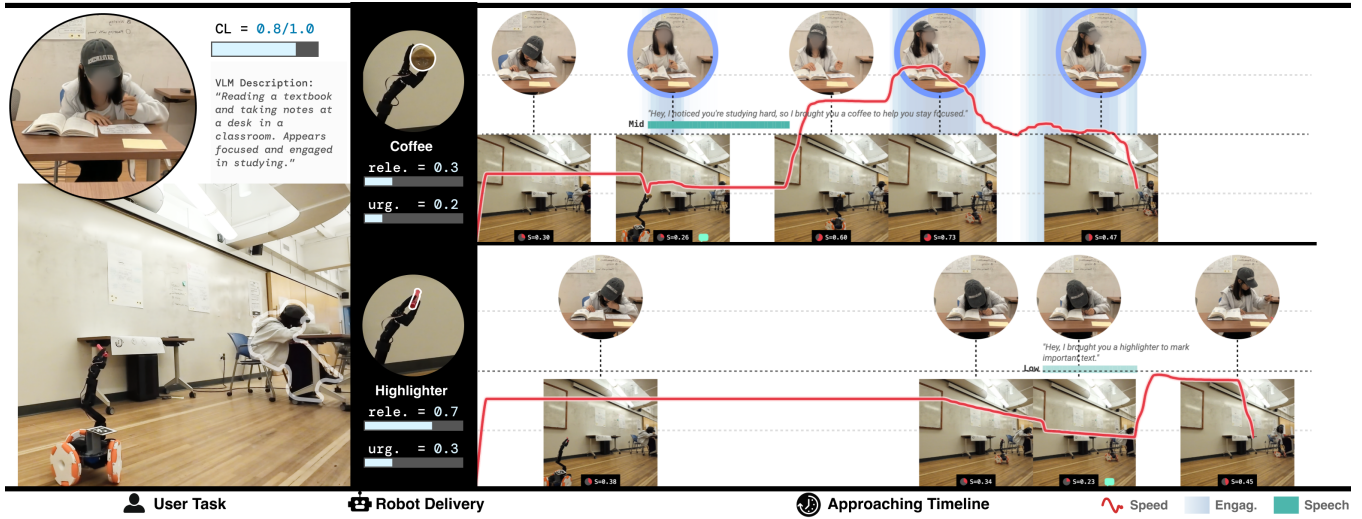


Figure 10: PROACH adapts delivery behavior to the user’s state. When the user is receptive (top), the robot speaks sooner and advances quickly for the coffee. When the user remains focused (bottom), the robot defers speech and approaches slowly for the highlighter.

Maya is reading a textbook and jotting notes at her desk late in the evening ($CL = 0.8$). The robot is meant to deliver a highlighter to support her note-taking, but detects her sustained focus and holds back: it approaches slowly, stays quiet, and only upon reaching her desk offers a brief reminder. Maya registers it peripherally, finishes her current thought, and reaches for the highlighter without breaking her rhythm.

If instead Maya happens to be more receptive when the robot arrives with a coffee, having just set down her pen and glanced up, it detects the microbreak and advances more quickly. By the time it speaks, “Hey, I noticed you’re studying hard, so I brought you a coffee to help you stay focused,” Maya is already attending; she takes the cup without comment.

5.3 Supporting Mixed-Initiative Interaction

By adapting approach speed and explanation timing to context, PROACH naturally preserves response windows: a slower approach gives the user time to decline before the handoff completes, while a faster approach draws attention early enough for the user to interrupt and accept (Fig. 11).

Lily is painting on a canvas in her living room ($CL = 0.2$) with moderate engagement but a leisurely pace. Her phone buzzed silently on the dining room table. When the robot moves to deliver the phone, it closes the distance quickly, which attracts her attention. As she glances up, the robot begins to explain while slowing down slightly. She recognizes the situation and interrupts before it finishes: “Yes, bring it to me.” The robot stops speaking, accelerates, hands over the phone, and retreats.

If instead the robot is delivering a *cushion*, its behavior differs: it approaches at a moderate pace. When it reaches conversational range, it explains: “Hey, I noticed you’re painting on the couch and brought you a cushion. It should make your...” This gives Lily a moment to consider, and she declines: “No thanks.” The robot turns around and leaves.

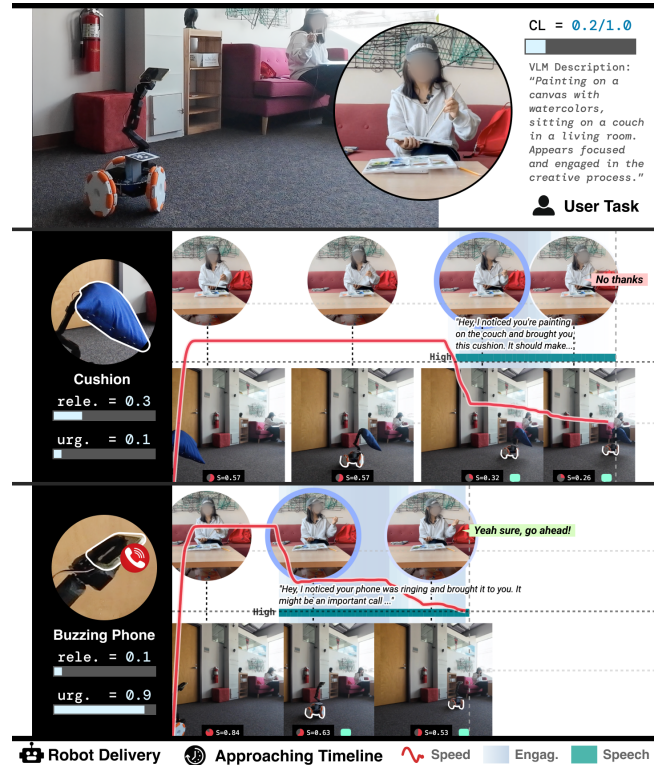


Figure 11: PROACH preserves response windows for mixed-initiated interaction. A buzzing phone triggers a fast approach that the user intercepts; a cushion yields a slower approach with more explanation, giving time to decline.

6 DISCUSSION

Design recommendations. We derive design considerations for proactive robot behavior from our findings:

- (1) **Speech matters.** A silent robot with context-aware speed modulation still leaves users confused. And the adaptive speech content further improves clarity: explanation detail should scale inversely with how self-evident the delivery is, as task-relevant items need minimal justification while externally motivated ones require explicit context.
- (2) **Adaptive speed changes should be paired with communicative context;** without it, speed variation can feel unpredictable rather than responsive.
- (3) Designers should anticipate that **users may interpret functional adaptations as social signals**, even without an explicit affective model, and that the alignment of speech and movement produces stronger social impressions than either channel alone.
- (4) **Stable individual differences** in speech timing preference suggest that personalization of when to speak may matter more than optimizing timing from task context alone.

Limitations in evaluation. Several limitations constrain the evaluation of *PROACH*. First, the silent, fixed-speed baseline may amplify the apparent benefit of speech, despite our efforts to bridge the large gap between the current state of the art and our multimodal system by introducing the ablated versions *BaSpeak* and *OptSpeed*. Second, the evaluation relies mainly on subjective ratings, and future work should measure objective task disruption, including efficiency, attention recovery, and task continuity. Third, the controlled setup used a fixed right-side, straight-line approach, excluding dynamic obstacles and varied approach angles. Finally, the small, demographically narrow sample ($N=12$) may limit statistical power. This may explain the non-significant effects of movement style and vocal timing, although qualitative data indicates that *PROACH*'s joint formulation is valuable to users.

Toward richer adaptation and deployment. To isolate the performance of the core optimizer in our controlled user study, the context variables and users' cognitive load were predefined, and the triggering of each delivery was manually controlled. Yet, we envision an end-to-end delivery pipeline: (1) background user activity assessment, (2) an automated trigger based on inferred user needs, (3) proactive item selection and retrieval, (4) real-time approach optimization, (5) interruption execution, and (6) response handling. To close the loop, our exploratory VLM component serves as a scalable demonstration of the full pipeline. It illustrates how real-time video input can be utilized to automate the upstream stages, including inferring task context and assessing cognitive load, before being handed over to the real-time approach optimization layer.

PROACH's middleware architecture is designed to integrate advanced sensory, environmental, and interactive layers seamlessly in the future without altering our three core terms (disturbance, delay, clarity) via two distinct mechanisms:

- (1) **Dynamic Parameter Adaptation:** Future work could extend *PROACH* by adapting the currently fixed cost weights

and thresholds to individual users over time. This would support personalization by learning latent preferences or incorporating explicit preferences as hard constraints. It could also enable multi-turn interaction by adjust the policy across successive attempts when a user misses, ignores, or does not respond to an initial cue.

- (2) **Modular Cost Integration:** New sensing signals, perception models, and output modalities could be incorporated into the three core objectives as additional sub-terms or constraints. For example, ambient noise could influence speech volume, while obstacle proximity could guide non-straight approach trajectories. The framework could also incorporate more expressive robot behaviors, including hesitation gestures and speech tone, when on-the-fly expressive behaviors more accessible in recent works [44]. In each case, the core objectives of disturbance, delay, and clarity would remain unchanged; only the action space and its mapping to these objectives would need to be extended.

7 CONCLUSION

The main contribution of this work is a system that bridges contextual considerations and on-robot approaching behavior, balancing clarity, disturbance, and efficiency for helpful but not distracting robot delivery. We evaluated our system through a user study and demonstrated initial feasibility of a perceive–reason–actuate loop for proactive robot delivery in real-world environments in our application examples. Future work should further close the loop with user-specific learning, richer context sensing, and additional output modalities.

ACKNOWLEDGMENTS

We thank Yujia Liu, Yitao Fan, David Lindlbauer, and Yi-Fei Cheng for insightful discussions. We thank Violet Han, Daehwa Kim, Yi-Hao Peng, and Yulin Chen for helping with the study setup.

This work was partially supported by the National Science Foundation under Award No. IIS-2443190.

REFERENCES

- [1] Riku Arakawa, Hiromu Yakura, and Mayank Goel. 2024. PrISM-Observer: Intervention Agent to Help Users Perform Everyday Procedures Sensed using a Smartwatch. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) (UIST '24). Association for Computing Machinery, New York, NY, USA, Article 15, 16 pages. doi:10.1145/3654777.3676350
- [2] Kilian Bahnsen, Linn Tiemann, Leonhard Plabst, and Tobias Grundgeiger. 2024. Augmented Reality Cues Facilitate Task Resumption after Interruptions in Computer-Based and Physical Tasks. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (CHI '24). Association for Computing Machinery, Honolulu, HI, USA, Article 273, 16 pages. doi:10.1145/3613904.3642666
- [3] Brian P. Bailey, Joseph A. Konstan, and John V. Carlis. 2001. The Effects of Interruptions on Task Performance, Annoyance, and Anxiety in the User Interface. In *Proceedings of INTERACT*. 593–601.
- [4] Saskia Bakker, Elise van den Hoven, and Berry Eggen. 2015. Peripheral Interaction: Characteristics and Considerations. *Personal and Ubiquitous Computing* 19, 1 (2015), 239–254. doi:10.1007/s00779-014-0775-2
- [5] Siddhartha Banerjee and Sonia Chernova. 2017. Temporal Models for Robot Classification of Human Interruptibility. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*. IFAAMAS, 1350–1359.
- [6] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* 1, 1 (2009), 71–81. doi:10.1007/s12369-008-0001-3

- [7] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. doi:10.18637/jss.v067.i01
- [8] Simon J. Blanchard and Ishani Banerji. 2016. Evidence-based recommendations for designing free-sorting experiments. *Behavior Research Methods* 48, 4 (2016), 1318–1336. doi:10.3758/s13428-015-0644-6
- [9] Dražen Bršćić, Takayuki Kanda, Tetsushi Ikeda, and Takahiro Miyashita. 2013. Person Tracking in Large Public Spaces Using 3-D Range Sensors. *IEEE Transactions on Human-Machine Systems* 43, 6 (2013), 522–534. doi:10.1109/THMS.2013.2283945
- [10] Sera Buyukgoz, Jasmin Grosinger, Mohamed Chetouani, and Alessandro Saffioti. 2022. Two Ways to Make Your Robot Proactive: Reasoning about Human Intentions, or Reasoning about Possible Futures. *Frontiers in Robotics and AI* 9, Article 929267 (2022). doi:10.3389/frobt.2022.929267
- [11] Remi Cadene, Simon Alibert, Alexander Sadeghi, Eduard Fons, et al. 2024. LeRobot: State-of-the-art Machine Learning for Real-World Robotics in PyTorch. <https://github.com/huggingface/lerobot>.
- [12] Ruei-Che Chang, Yuxuan Liu, and Anhong Guo. 2024. WorldScribe: Towards Context-Aware Live Visual Descriptions. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3654777.3676375
- [13] Yifei Cheng, Yukang Yan, Xin Yi, Yuanchun Shi, and David Lindlbauer. 2021. SemanticAdapt: Optimization-based Adaptation of Mixed Reality Layouts Leveraging Virtual-Physical Semantic Connections. In *Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology (UIST)*. ACM, 282–297. doi:10.1145/3472749.3474750
- [14] Yi-Shan Chiang, Tsung-Shien Chu, Chyi Dong Lim, Tsu-Yun Wu, Shao-Hua Tseng, and Li-Chen Fu. 2014. Personalizing Robot Behavior for Interruption in Social Human-Robot Interaction. In *Proceedings of the IEEE International Workshop on Advanced Robotics and its Social Impacts*. IEEE, 44–49.
- [15] Rune Haubo B Christensen. 2018. *Cumulative link models for ordinal regression with the R package ordinal*. R package vignette.
- [16] Kerstin Dautenhahn, Michael Walters, Sarah Woods, Kheng Lee Koay, Christopher L Nehaniv, A Sisbot, Rachid Alami, and Thierry Siméon. 2006. How may I serve you? A robot companion approaching a seated person in a helping context. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. 172–179.
- [17] Francesco Del Duchetto, Paul Baxter, and Marc Hanheide. 2020. Are You Still with Me? Continuous Engagement Assessment from a Robot's Point of View. *Frontiers in Robotics and AI* 7 (2020). doi:10.3389/frobt.2020.00116
- [18] Gonzalo Ferrer, Anais Garrell, and Alberto Sanfeliu. 2013. Robot Companion: A Social-Force Based Approach with Human Awareness-Navigation in Crowded Environments. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Tokyo, Japan, 1688–1694. doi:10.1109/IROS.2013.6696577
- [19] Krzysztof Gajos and Daniel S. Weld. 2004. SUPPLE: Automatically Generating User Interfaces. In *Proceedings of the 9th International Conference on Intelligent User Interfaces (IUI)*. ACM, 93–100. doi:10.1145/964442.964461
- [20] Krzysztof Z. Gajos, Daniel S. Weld, and Jacob O. Wobbrock. 2010. Automatically Generating Personalized User Interfaces with Supple. *Artificial Intelligence* 174, 12–13 (2010), 910–950. doi:10.1016/j.artint.2010.05.005
- [21] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. 2020. Towards Fast, Accurate and Stable 3D Dense Face Alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- [22] Edward T. Hall. 1966. *The Hidden Dimension*. Doubleday, New York, NY, USA.
- [23] Violet Yinuo Han, Jesse T Gonzalez, Christina Yang, Zhiruo Wang, Scott E Hudson, and Alexandra Ion. 2025. Towards Unobtrusive Physical AI: Augmenting Everyday Objects with Intelligence and Robotic Movement for Proactive Assistance. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 173, 16 pages. doi:10.1145/3746059.3747726
- [24] N. Hanajima, T. Goto, Y. Ohta, H. Hikita, and M. Yamashita. 2005. A motion rule for human-friendly robots based on electrodermal activity investigations and its application to mobile robot. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 3791–3797. doi:10.1109/IROS.2005.1545301
- [25] Dirk Helbing and Péter Molnár. 1995. Social Force Model for Pedestrian Dynamics. *Physical Review E* 51, 5 (1995), 4282–4286. doi:10.1103/PhysRevE.51.4282
- [26] Zachary Henkel, Cindy L Bethel, Robin Roberson Murphy, and Vasant Srinivasan. 2014. Evaluation of proxemic scaling functions for social robotics. *IEEE Transactions on Human-Machine Systems* 44, 3 (2014), 374–385.
- [27] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 159–166. doi:10.1145/302979.303030
- [28] Eric Horvitz and Johnson Apacible. 2003. Learning and Reasoning about Interruption. In *Proceedings of the 5th International Conference on Multimodal Interfaces (ICMI)*. ACM, 20–27. doi:10.1145/958432.958440
- [29] Eric Horvitz, Andy Jacobs, and David Hovel. 1999. Attention-Sensitive Alerting. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI)*. Morgan Kaufmann, 305–313.
- [30] Eric Horvitz, Paul Koch, Raman Sarin, Johnson Apacible, and Muru Subramani. 2004. BusyBody: Creating and Fielding Personalized Models of the Cost of Interruption. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work (CSCW)*. ACM, 507–510. doi:10.1145/1031607.1031690
- [31] Shamsi T. Iqbal and Brian P. Bailey. 2006. Leveraging Characteristics of Task Structure to Predict Costs of Interruption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 741–750. doi:10.1145/1124772.1124882
- [32] Mitsuhiro Kamezaki, Haoyi Yang, Hiroyasu Iwata, and Shigeki Sugano. 2019. A Framework for Autonomous Search-and-Rescue Robots Using Behavior Trees and Human-Robot Interaction. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. Montreal, QC, Canada, 6418–6424. doi:10.1109/ICRA.2019.8794178
- [33] Yusuke Kato, Takayuki Kanda, and Hiroshi Ishiguro. 2015. May I help you? Design of Human-like Polite Approaching Behavior. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (Portland, Oregon, USA) (HRI '15)*. Association for Computing Machinery, New York, NY, USA, 35–42. doi:10.1145/2696454.2696463
- [34] Jens Kessler, Christof Schroeter, and Horst-Michael Gross. 2011. Approaching a person in a socially acceptable manner using a fast marching planner. In *Proceedings of the 4th International Conference on Intelligent Robotics and Applications - Volume Part II (Aachen, Germany) (ICIRA'11)*. Springer-Verlag, Berlin, Heidelberg, 368–377. doi:10.1007/978-3-642-25489-5_36 approach: just traj. planning.
- [35] Marike Koch van den Broek and Thomas B. Moeslund. 2024. What is Proactive Human-Robot Interaction? A Review of a Progressive Field and Its Definitions. *ACM Transactions on Human-Robot Interaction* 13, 4 (2024), 1–30. doi:10.1145/3643803
- [36] Geonsun Lee, Min Xia, Nels Numan, Xun Qian, David Li, Yanhe Chen, Achin Kulshrestha, Ishan Chatterjee, Yinda Zhang, Dinesh Manocha, David Kim, and Ruofei Du. 2025. Sensible Agent: A Framework for Unobtrusive Interaction with Proactive AR Agents. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 119, 22 pages. doi:10.1145/3746059.3747748
- [37] Russell V. Lenth. 2024. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. <https://CRAN.R-project.org/package=emmeans> R package version 1.10.0.
- [38] Chenyi Li, Guande Wu, Gromit Yeuk-Yin Chan, Dishita Gdi Turakhia, Sonia Castelo Quispe, Dong Li, Leslie Welch, Claudio Silva, and Jing Qian. 2025. Satori: Towards Proactive AR Assistant with Belief-Desire-Intention User Modeling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1229, 24 pages. doi:10.1145/3706598.3714188
- [39] Jiahao Nick Li, Yan Xu, Tovi Grossman, Stephanie Santosa, and Michelle Li. 2024. OmniActions: Predicting Digital Actions in Response to Real-World Multimodal Sensory Inputs with LLMs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM. doi:10.1145/3613904.3642068
- [40] David Lindlbauer, Anna Maria Feit, and Otmár Hilliges. 2019. Context-Aware Online Adaptation of Mixed Reality Interfaces. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (UIST)*. ACM, 147–160. doi:10.1145/3332165.3347945
- [41] Huaping Liu, Di Guo, and Angelo Cangelosi. 2025. Embodied Intelligence: A Synergy of Morphology, Action, Perception and Learning. *Comput. Surveys* 57, 7, Article 186 (2025), 36 pages. doi:10.1145/3717059
- [42] Tianjian Liu, Hongzheng Zhao, Yuheng Liu, Xingbo Wang, and Zhenhui Peng. 2024. ComPeer: A Generative Conversational Agent for Proactive Peer Support. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (Pittsburgh, PA, USA) (UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 117, 22 pages. doi:10.1145/3654777.3676430
- [43] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. 2025. Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI. *IEEE/ASME Transactions on Mechatronics* 30 (2025), 7253–7274.
- [44] Karthik Mahadevan, Jonathan Chien, Noah Brown, Zhuo Xu, Carolina Parada, Fei Xia, Andy Zeng, Leila Takayama, and Dorsa Sadigh. 2024. Generative Expressive Robot Behaviors using Large Language Models. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (Boulder, CO, USA) (HRI '24)*. Association for Computing Machinery, New York, NY, USA, 482–491. doi:10.1145/3610977.3634999
- [45] Swathi Mannem, William Macke, Peter Stone, and Reuth Mirsky. 2023. Exploring the Cost of Interruptions in Human-Robot Teaming. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots (Humanoids)*. IEEE. doi:10.1109/Humanoids57100.2023.10375189
- [46] Daniel C. McFarlane and Kara A. Latorella. 2002. The Scope and Importance of Human Interruption in Human-Computer Interaction Design. 17, 1 (2002), 1–61. doi:10.1207/S15327051HCI11701_1

- [47] Ross Mead and Maja J. Matarić. 2017. Autonomous human–robot proxemics: socially aware navigation based on interaction potential. *Auton. Robots* 41, 5 (June 2017), 1189–1201. doi:10.1007/s10514-016-9572-2
- [48] M.P. Michalowski, S. Sabanovic, and R. Simmons. 2006. A spatial model of engagement for a social robot. In *9th IEEE International Workshop on Advanced Motion Control, 2006*. 762–767. doi:10.1109/AMC.2006.1631755
- [49] Jonathan Mumm and Bilge Mutlu. 2011. Human-Robot Proxemics: Physical and Psychological Distancing in Human-Robot Interaction. In *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM, 331–338. doi:10.1145/1957656.1957786
- [50] Richard W. Obermayer and William A. Nugent. 2000. Human-Computer Interaction for Alert Warning and Attention Allocation Systems of the Multimodal Watchstation. In *International Symposium on Optical Science and Technology*, Patricia Hamburger (Ed.). San Diego, CA, 14–22. doi:10.1117/12.407536
- [51] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*. ACM, 1–22. doi:10.1145/3586183.3606763
- [52] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galiker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haochuan Wang, Lili Yu, and Ury Zhilinsky. 2025. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. *arXiv preprint arXiv:2504.16054* (2025).
- [53] Karola Pitsch, Hideaki Kuzuoka, Yuya Suzuki, Luke Sussenbach, Paul Luff, and Christian Heath. 2009. "The First Five Seconds": Contingent Stepwise Entry into an Interaction as a Means to Secure Sustained Engagement in HRI. In *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 985–991. doi:10.1109/ROMAN.2009.5326167
- [54] Kevin Pu, Ting Zhang, Naveen Sendhilnathan, Sebastian Freitag, Raj Sodhi, and Tanya R. Jonker. 2025. ProMemAssist: Exploring Timely Proactive Assistance Through Working Memory Modeling in Multi-Modal Wearable Devices. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 56, 19 pages. doi:10.1145/3746059.3747770
- [55] Vincent Rousseau, François Ferland, Dominic Létourneau, and François Michaud. 2013. Sorry to interrupt, but may I have your attention?: preliminary design and evaluation of autonomous engagement in HRI. *J. Hum. Robot Interact.* 2, 3 (2013), 41–61.
- [56] S. M. Bhagya P. Samarakoon, M. A. Viraj J. Muthugala, and A. G. Buddhika P. Jayasekara. 2022. A Review on Human–Robot Proxemics. *Electronics* 11, 16 (2022), 2490. doi:10.3390/electronics11162490
- [57] Satoru Satake, Takayuki Kanda, Dylan F. Glas, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. 2009. How to Approach Humans? Strategies for Social Robots to Initiate Interaction. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. La Jolla, CA, USA, 109–116. doi:10.1145/1514095.1514118
- [58] Paul Saulnier, Ehud Sharlin, and Saul Greenberg. 2010. Exploring minimal nonverbal interruption in social hri. (2010).
- [59] Chao Shi, Satoru Satake, Takayuki Kanda, and Hiroshi Ishiguro. 2017. A Robot That Distributes Flyers to Pedestrians in a Shopping Mall. *International Journal of Social Robotics* 10, 4 (Nov. 2017), 421–437. doi:10.1007/s12369-017-0442-7
- [60] Zhi Shi, Thomas Groechel, and Maja J. Matarić. 2018. Proxemics-Based Adaptive Shared Control for Human–Robot Collaborative Navigation. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. Nanjing, China, 1030–1035. doi:10.1109/ROMAN.2018.8525547
- [61] Ben Shneiderman and Pattie Maes. 1997. Direct manipulation vs. interface agents. *Interactions* 4, 6 (Nov. 1997), 42–61. doi:10.1145/267505.267514
- [62] Candace Sidner, Christopher Lee, Cory Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence* 166 (08 2005), 140–164. doi:10.1016/j.artint.2005.03.005
- [63] John Sweller. 2011. Cognitive Load Theory. In *Psychology of Learning and Motivation*. Vol. 55. Academic Press, 37–76.
- [64] Leila Takayama and Caroline Pantofaru. 2009. Influences on Proxemic Behaviors in Human-Robot Interaction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 5495–5502. doi:10.1109/IROS.2009.5354145
- [65] Kazunori Terada, Takashi Shamoto, Haiying Mei, and Akira Ito. 2007. Reactive Movements of Non-Humanoid Robots Cause Intention Attribution in Humans. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 3715–3720. doi:10.1109/IROS.2007.4399429
- [66] Vaibhav V. Unhelkar, Shen Li, and Julie A. Shah. 2020. Decision-Making for Bidirectional Communication in Sequential Human-Robot Collaborative Tasks. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM, 329–341.
- [67] Marlieke T. R. Van Kesteren, Dirk J. Ruiter, Guillén Fernández, and Richard N. Henson. 2012. How Schema and Novelty Augment Memory Formation. *Trends in Neurosciences* 35, 4 (2012), 211–219.
- [68] Brandon Victor Syiem and Eduardo Velloso. 2026. Better Assumptions, Stronger Conclusions: The Case for Ordinal Regression in HCI. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 820, 21 pages. doi:10.1145/3772318.3790821
- [69] Michael L. Walters, Kerstin Dautenhahn, Rik te Boekhorst, Kheng Lee Koay, Dag Syrdal, and Chrystopher L. Nehaniv. 2009. An Empirical Framework for Human-Robot Proxemics. In *Proceedings of New Frontiers in Human-Robot Interaction*. 144–149.
- [70] Zeyi Wang, Yuqian Bansal, Yifan Tao, Chien-Ming Wu, and Julie Shah. 2024. Mixed-Initiative Human-Robot Teaming under Suboptimality with Online Bayesian Adaptation. In *Proceedings of the 23rd International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*. IFAAMAS.
- [71] Chang Xiao and Zixiaofan Yang. 2025. Streaming, Fast and Slow: Cognitive Load-Aware Streaming for Efficient LLM Serving. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 71, 13 pages. doi:10.1145/3746059.3747721
- [72] Haifeng Xu, Chunwen Li, Xiaohu Yuan, Tao Zhi, and Huaping Liu. 2025. A simulated dataset for proactive robot task inference from streaming natural language dialogues. *Scientific Data* 12, 1 (2025), 1405. doi:10.1038/s41597-025-05727-w
- [73] Yuto Yamaji, Taisuke Miyake, Yuta Yoshiike, P. Ravindra S. De Silva, and Michio Okada. 2010. STB: Intentional Stance Grounded Child-Dependent Robot. In *Proceedings of the International Conference on Social Robotics (ICSR)*. Springer, 69–78. doi:10.1007/978-3-642-17248-9_8
- [74] Takuma Yoneda, Matthew Black, Joonhyung Huh, Matthew R. Walter, and Maani Ramanathan. 2025. Agreeing to Interact in Human-Robot Interaction using Large Language Models and Vision Language Models. *arXiv preprint arXiv:2503.15491* (2025).