

Assist-As-Needed: Adaptive Multimodal Robotic Assistance for Medication Management in Dementia Care

Kruthika Gangaraju
kgangaraju@wpi.edu
Worcester Polytechnic Institute
Worcester, Massachusetts, USA

Tanmayi Inaparthi
tinaparthi@wpi.edu
Worcester Polytechnic Institute
Worcester, Massachusetts, USA

Jiaqi Yang
jyang6@wpi.edu
Worcester Polytechnic Institute
Worcester, Massachusetts, USA

Yihao Zheng
yzheng8@wpi.edu
Worcester Polytechnic Institute
Worcester, Massachusetts, USA

Fengpei Yuan
fyuan3@wpi.edu
Worcester Polytechnic Institute
Worcester, Massachusetts, USA

Abstract

People living with dementia (PLWDs) face progressively declining abilities in medication management—from simple forgetfulness to complete task breakdown—yet most assistive technologies fail to adapt to these changing needs. This one-size-fits-all approach undermines autonomy, accelerates dependence, and increases caregiver burden. Occupational therapy principles emphasize matching assistance levels to individual capabilities: minimal reminders for those who merely forget, spatial guidance for those who misplace items, and comprehensive multimodal support for those requiring step-by-step instruction. However, existing robotic systems lack this adaptive, graduated response framework essential for maintaining PLWD independence. We present an adaptive multimodal robotic framework using the Pepper robot that dynamically adjusts assistance based on real-time assessment of user needs. Our system implements a hierarchical intervention model progressing from (1) simple verbal reminders, to (2) verbal + gestural cues, to (3) full multimodal guidance combining physical navigation to medication locations with step-by-step verbal and gestural instructions. Powered by LLM-driven interaction strategies and multimodal sensing (verbal, touch-screen, and camera), the system continuously evaluates task states to provide just-enough assistance—preserving autonomy while ensuring medication adherence. We conducted a preliminary study with healthy adults and dementia care stakeholders in a controlled lab setting, evaluating the system’s usability, comprehensibility, and appropriateness of adaptive feedback mechanisms. This work contributes: (1) a theoretically-grounded adaptive assistance framework translating occupational therapy principles into HRI design, (2) a multimodal robotic implementation that preserves PLWD dignity through graduated support, and (3) empirical insights into stakeholder perceptions of adaptive robotic care. Our findings establish foundations for future trials with PLWDs, advancing toward truly person-centered robotic dementia care that augments rather than replaces human caregiving.

Keywords

Socially assistive robotics, Multimodal human-robot interaction, Robot-assisted dementia care

1 Introduction

The population of adults aged 65 and above has increased rapidly with rising life expectancy globally. In the United States alone, there

were an estimated 54.1 million older adults in 2019, a number projected to exceed 80 million by 2040 [8]. As adults 65 and older now comprise roughly one in seven Americans, age-related disorders have become more prevalent. Cognitive decline occurs due to aging [33], and when it gets severe enough to affect independence in activities of daily living (ADLs), it is classified as dementia [24]. Approximately 10 million new dementia cases are reported worldwide each year, with Alzheimer’s disease (AD)—the most common cause—accounting for roughly 60–80% of cases [7]. In the United States, an estimated 6.9 million adults aged 65+ were living with AD or AD-related dementias (ADRD) in 2024. AD/ADRD are progressive neurodegenerative conditions characterized by cognitive, behavioral, and/or motor impairments [34] that reduce independence and quality of life [1]. People living with dementia experience declines in memory, language, and problem-solving that hinder activities of daily living (ADLs), often necessitating assistance to initiate and complete tasks [1]. One of the commonly faced issue by PLWDs is progressively lose the ability to manage medications, from minor forgetfulness to complete task breakdown [2], which in turn imposes significant strain on the caregiver. Although no cure exists to date, research increasingly targets early diagnostic approaches and supportive interventions. Among these, artificial intelligence (AI) and socially assistive robotics (SAR) have emerged to support cognitive, emotional, and physical aspects of care for individuals with AD/ADRD [16], with the potential to reduce caregiver burden and to detect and mitigate ADL difficulties.

SAR is effective in providing the necessary reminders and guidance required for dementia caregivers. While it is effective, it also requires powerful, reliable, and trustworthy AI that can identify forgetful moments and deliver appropriate, timely assistance at the right time and place [15, 36]. The goal of having SAR for caregiving is to prompt the PLWD to initiate and complete their task by providing *minimal yet effective* support without taking away their independence, which may contribute to slower decline of cognitive abilities and disease progression [26]. While early robotic assistants such as [12, 29] and more recent systems such as RAMCIP demonstrated that mobile robots can issue reminders and even physically deliver objects [21], these systems largely provide fixed, one-size-fits-all interventions. They do not systematically adjust assistance to a person’s fluctuating cognitive and functional abilities, nor do they operationalize the **assist-as-needed** principles long

established in occupational therapy. This gap limits their capacity to preserve autonomy and dignity as dementia progresses.

Additionally, existing robotic systems for dementia care typically rely on bespoke speech recognition and rule-based interaction flows, restricting their ability to handle ambiguous or unexpected human behaviors [35]. Even when speech, vision, and navigation are present, modalities are often stitched together with brittle heuristics rather than a unified, uncertainty-aware decision layer, making cross-modal conflicts and sensor dropouts hard to manage. They also rarely involve care professionals or stakeholders early in the design process, leaving uncertainty about acceptability and appropriateness. We address these limitations by introducing one of

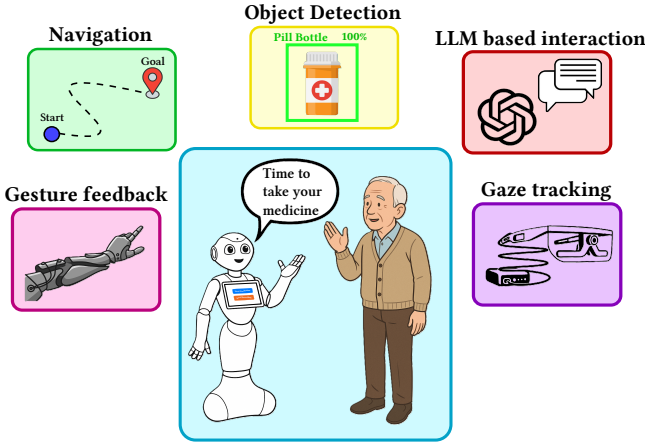


Figure 1: Adaptive Assist-As-Needed for people with dementia, involving multimodal perception and feedback

the first robotic system to explicitly translate occupational therapy’s adaptive “assist-as-needed” framework as a concrete, testable human–robot interaction pipeline for medication management in dementia care. Running on the Pepper platform, our system combines GPT-4o [17] for dialogue and task reasoning, Whisper [32] for speech recognition, and YOLOv11 [18] for visual understanding of medication artifacts; it fuses these with gesture feedback, an on-device touchscreen, and autonomous navigation (Figure 1). This approach preserves autonomy and dignity, while ensuring support when needed. We also evaluate the system with healthy adults and dementia care stakeholders to validate usability, comprehensibility, and appropriateness of adaptive feedback mechanisms.

Guided by the goal of “HRI Empowering Society”, supporting independence, dignity, and equitable access to care, we investigate two research questions: **RQ1)** How effective is the involvement of a social robot in supporting medication adherence relative to a verbal-only baseline? **RQ2)** What level of assistance (e.g., minimal cues vs. step-by-step guidance) do users prefer, and how does it affect perceived workload and autonomy? To answer these questions, we compare our adaptive, multimodal system against a *verbal-only* baseline that offers reminders and reassurance but provides no step-wise instruction, on-device visuals, pointing/gestures, or shared control. We operationalize adherence via task completion, complement these with efficiency measures (time on task; number of interactions), and assess experience using standardized ratings of

workload and perceived autonomy. Our a priori expectation is that the adaptive, embodied approach will increase adherence and reduce workload *without* undermining autonomy—and may improve it by offering the *least assistance necessary* at each step.

In doing so, our work speaks to empowerment at three levels: (i) for people living with dementia, by preserving agency through graded, user-controlled assistance; (ii) for caregivers and clinicians, by making assistance transparent, configurable, and reviewable; and (iii) for health systems, by demonstrating an interpretable decision policy that can generalize beyond scripted lab tasks. We report quantitative and qualitative outcomes from healthy adults and dementia-care stakeholders to evaluate usability, comprehensibility, and the appropriateness of adaptive feedback, and to surface design principles for dignifying, autonomy-preserving HRI in medication management.

2 Related Work

Prior work has explored assistive robots as companions for people living with dementia (PLWDs) [19, 22], examined the acceptance of social robots in care settings [20], and investigated their use in rehabilitative therapy [6]. Despite this progress, usability and effectiveness for concrete assistance with activities of daily living (ADLs) in real-world settings remain under-examined [11].

Within dementia care, robots have been deployed for task-level prompting and graded assistance. Wang et al. [35] implemented a teleoperated socially assistive robot that delivered need-based assistance and reported acceptance from both PLWD and caregiver perspectives; however, the system depended on a human teleoperator rather than automated decision making. Mobile platforms have also demonstrated physical assistance: the Nursebot “Pearl” [29] integrated an actuated head, onboard sensors, speech synthesis, and a touchscreen to provide navigational support and time-contingent medication reminders for older adults, while Care-O-bot [12] explored manipulation for pick-and-place of common indoor objects to assist older adults and people with physical disabilities. Neither system centered social interaction or the principle of providing help only when needed to preserve PLWDs’ independence. More recently, Pepper has been used in dementia-relevant task contexts: Yuan et al. [37, 38] teleoperated Pepper to assist PLWDs with PASS tasks, examining assistance levels and underscoring the need for adaptive, personalized interactions.

Across this literature, a recurring theme is the integration of multimodal feedback to support smooth task execution while reducing cognitive load. Typical modalities include (i) auditory prompts (natural-language reminders, stepwise instructions, confirmations), (ii) visual cues on an integrated screen (checklists, icons, timing indicators), and (iii) embodied signals such as pointing, head orientation and gaze, and purposeful proxemics (navigating to the relevant location or object), with manipulation when available. Effective systems blend these channels to enable graded cueing—from high-level reminders to fine-grained step guidance—clarify environmental referents, and support error recovery, while preserving user autonomy and accommodating individual preferences.

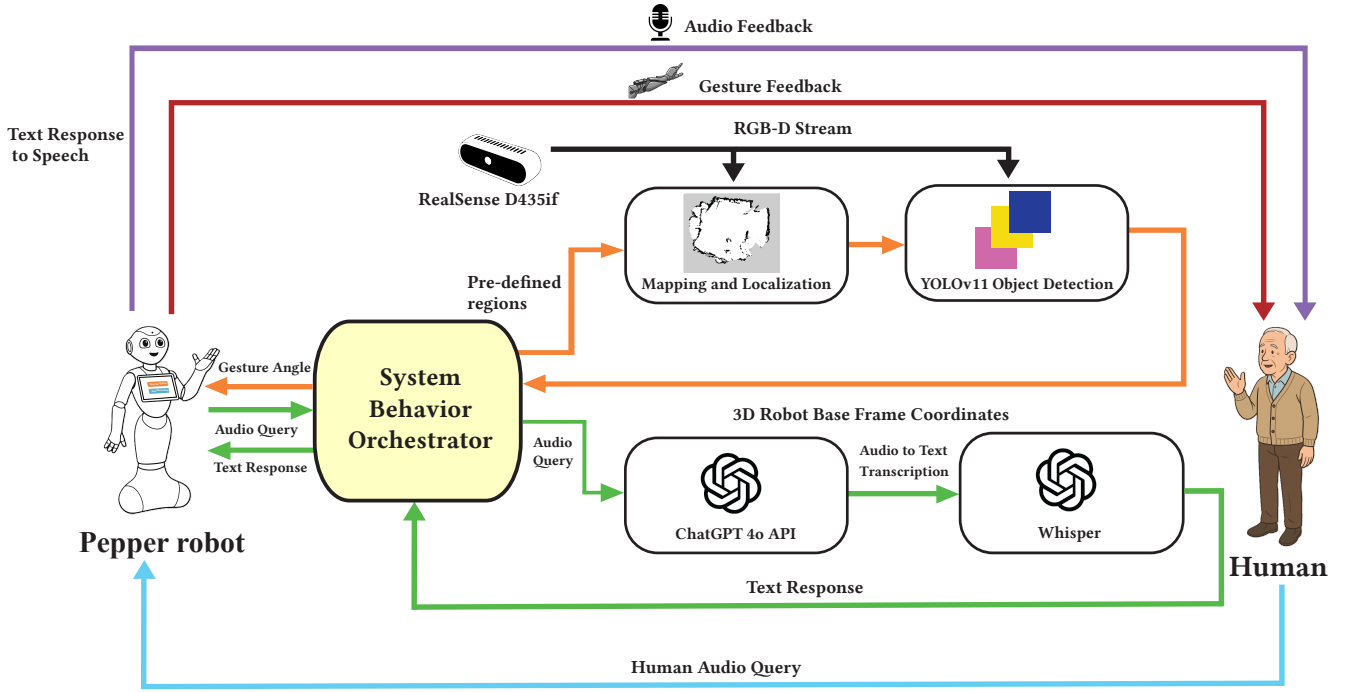


Figure 2: Overview of the framework for medication management. The arrows indicate the flow of information.

3 Methodology

We developed an adaptive multimodal robotic framework that integrates multimodal perception (audio, touchscreen, and camera/visual, and eye-tracking) with an LLM-backed dialogue manager to deliver adaptive, multimodal assistive guidance during a medication-taking routine for people with dementia. Implemented on the Pepper humanoid robot, the system combines verbal and gestural feedback - including gaze alignment and arm pointing - to support task completion while preserving autonomy. Figure 2 illustrates the system framework. In our implementation, the robot localizes against a pre-built map of the test environment and uses the standard ROS navigation stack to move to a designated region of interest (ROI). Regions of interest (ROIs) are predefined, manually labeled map locations where the pill bottle may be found. We selected this arbitrarily to simulate real-world scenario where the PLWDs could possibly leave their medication. Upon arrival to the ROI, the perception modules detect the pill bottle and water bottle, estimate the target’s 3-D pose, and generate embodied cues while concurrently delivering verbal and on-screen instructions. A system behavior orchestrator ensures the correct flow of information during the entire process. The navigation and object-detection stack was designed with complex deployment settings in mind (e.g., home environments or assisted-living centers). In such cases, multiple ROIs can be defined; the robot can visit each as needed, check for the pill bottle, and then guide the user to the exact location for the next steps. The system also incorporates an eye-tracking sensor to capture participants’ gaze patterns; although the full adaptive logic (e.g., prompting when prolonged gaze is detected) has not yet

been activated, this capability was used to collect gaze data from participants to inform future development.

3.1 Enhanced Robotic Platform with Integrated Sensors

In order to implement our framework, we made use of a SAR. Pepper robot is a widely used social robot in healthcare and hospitality sectors. The Pepper robot stands at 1.21-m tall and features 17 joints for graceful and expressive body movements. Pepper supports expressive non-verbal communication (gestures, posture, head gaze) and speech synthesis, and integrates RGB cameras, microphones, sonars/IR proximity sensors, and an on-chest tablet for visual prompts. Although Pepper is suitable for immersive, engaging human-robot interaction, Pepper’s native perception stack and depth sensing capabilities are limited for robust navigation and object detection. Additionally, its onboard computational power is insufficient for time-critical assistance and interaction with people with dementia. To address these limitations, we added an NVIDIA Jetson Orin NX module on its backplate for onboard inference and mounted an Intel RealSense D435i camera on top of its head, (as shown in Figure 3). The RealSense camera provides RGB-D streams, and the Orin runs SLAM preprocessing and object detection at the edge. Pepper’s on-board computer communicates with the Orin over Wi-Fi using ROS and a flask server.

3.2 Navigational Guidance

People living with dementia (PLWDs) may forget where medications or related items (e.g., water bottles) are placed, so the system autonomously visits likely storage locations to ground subsequent detection and guidance. We implement 2-D SLAM with

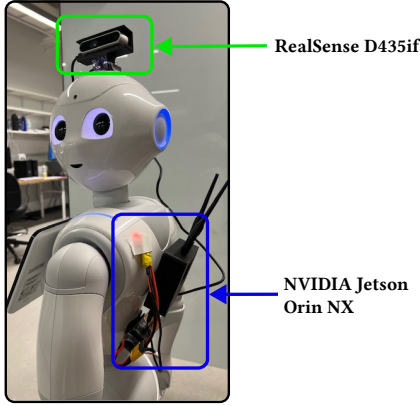


Figure 3: Hardware components mounted on Pepper robot

`slam_gmapping`, using the head-mounted RealSense depth stream converted to a virtual laser via `depthimage_to_laserscan`; an occupancy grid is built during an initial exploration pass [13]. For localization on the saved map, we use Adaptive Monte Carlo Localization (`amcl`), which maintains an online-adaptive particle filter over robot pose [9]. Path execution is managed by `move_base` [31]: the global planner (Dijkstra/A* over the static costmap) provides routes, and the local planner implements the Dynamic Window Approach to select feasible velocity commands [10]. Layered costmaps (`costmap_2d`) combine a static layer (map), an obstacle layer (virtual-laser/depth obstacles), and an inflation layer to maintain clearance. Navigation goals are specified as regions of interest (ROIs)—predefined map coordinates where the pill bottle may be found. At run time, the system behavior orchestrator (see Figure 2, orange line) sends ROI waypoints sequentially to `move_base`; Once the robot reaches an ROI, it pauses motion and triggers the YOLO detection module (see subsection 3.3). If the model does not detect the object, the robot moves to the next ROI. Velocity commands are published on `/cmd_vel` to the Pepper base controller, positioning the robot for the subsequent joint-attention and dialogue phases.

3.3 Multimodal Perception

Multimodal perception enables robots to operate in complex, human-centered settings by fusing heterogeneous signals—vision, speech / language, and touch—into a coherent model for interaction [25, 39]. In our system, camera-based perception is integrated with on-device touchscreen input and microphone audio to ground dialogue in the physical context.

Object Detection Module. After the robot localizes on a prebuilt map and navigates to the region of interest (ROI), it must reliably detect medication-related objects (pill bottle, water bottle) under variation in lighting, viewpoint, and clutter, while meeting real-time constraints on the Jetson. For these reasons, we use YOLOv11 [18] for object detection initialized from COCO weights [23]. We collected 1,200 images spanning diverse viewpoints, lighting conditions, and levels of clutter, and applied standard geometric and photometric augmentations to improve robustness. At the ROI, the system behavior orchestrator pans the head from -30° to $+30^\circ$ and acquires RGB frames at regular intervals; detection is run on

each frame, and upon a positive detection we extract the bounding box coordinates. To obtain a 3D target pose, we refine the detection to a per-pixel mask within the box using the aligned depth image: we compute the median depth Z_m inside the box, retain pixels whose depth lies within a narrow band around Z_m (depth-guided foreground extraction), and keep the largest connected component intersecting the box center as the segmented pill bottle. For each mask pixel (u, v) with depth $Z(u, v)$ and camera intrinsics (f_x, f_y, c_x, c_y) , we back-project to the camera frame using the pinhole model $X = \frac{(u-c_x)Z}{f_x}$, $Y = \frac{(v-c_y)Z}{f_y}$, $Z = Z$, stack $p_{\text{cam}} = [X, Y, Z, 1]^T$, and transform to the robot base via the calibrated extrinsics $p_{\text{base}} = T_{\text{base} \leftarrow \text{cam}} p_{\text{cam}}$. The resulting point cloud in the base frame is used to estimate a pointing direction: we take a small patch around the cloud centroid $\bar{p}$, fit a plane $ax+by+cz+d=0$ by least squares (normal $\mathbf{n}$ given by the eigenvector of the local covariance with the smallest eigenvalue), and orient the normal opposite the camera optical axis ($\mathbf{n} \cdot \hat{\mathbf{z}}_{\text{cam}} < 0$). Finally, we compute the pointing command from the vector $\mathbf{d} = \bar{p}_{\text{base}} - \mathbf{o}_{\text{arm}}$ to obtain yaw and pitch, $\psi = \text{atan2}(d_y, d_x)$ and $\theta = \text{atan2}(d_z, \sqrt{d_x^2 + d_y^2})$, which drive a brief arm gesture that directs user attention to the bottle (see Figure 2, orange data flow).

Audio Perception Module. The Pepper robot’s built-in microphones capture users’ spoken input. To initiate capture, the user presses the “Record” button presented on the robot’s tablet (push-to-talk). The resulting audio stream is passed to the Whisper model for automatic speech recognition (ASR), yielding a text transcript of the user’s response. This transcript is then forwarded to the downstream dialogue component (e.g., GPT-4o in our system) to interpret intent and generate the appropriate response. The audio module therefore provides a simple interface—explicit user-controlled recording, ASR transcription via Whisper, and text handoff to the interaction policy—so that higher-level modules can adapt assistance without relying on always-listening input (see Figure 2, blue and green data flow).

3.4 Multimodal Feedback

Motivated by prior HRI findings that complementary modalities improve comprehensibility and engagement [3, 27]—particularly for older adults and PLWD—we designed multimodal robotic feedback to minimize cognitive load and disambiguate references in the environment. Our design takes advantage of Pepper’s native features (speech synthesis, head / arm gestures, base orientation, and chest-mounted tablet) and delivers signals in two phases—verbal and deictic. In addition to these feedback modalities, the robot also provides *navigational feedback* by leading the user to the location of the pill bottle.

Gestural Feedback. Once the pill bottle is detected, the robot establishes joint attention by aligning head gaze with the 3D target and executing a brief, safe arm-pointing gesture (see Figure 2, red line). The robot maintains an appropriate standoff distance and, if necessary, repositions to ensure a clear line of sight.

Verbal Feedback. Throughout the task completion, the robot provides adaptive, detailed guidance with the steps the user has to follow (e.g., “Time to take your medicine, follow me!” / “Looking for

your medicine bottle"). The robot then provides concise, sequential instructions (e.g., "Open the bottle" → "Take the prescribed number of pills" → "Drink water"). Each step elicits an explicit confirmation or offers a repeat/rephrase on timeout.

Robot-user interaction is mediated by an LLM (GPT-4o) with a dialogue policy that balances task progress and safety to provide dementia-appropriate support during medication management. After navigating to the medication ROI, the robot verifies visual detection, delivers adaptive step-by-step instructions, and requests a final confirmation of intake before ending the routine (see Figure 2, purple line). The policy provides necessary guidance while minimizing distractions during execution. Prompts are engineered to handle edge cases (e.g., the user becomes distracted or expresses reluctance), to adapt tone and pacing to individual preferences, and to maintain a safe, supportive style that keeps users engaged and motivated to complete the task.

Across phases, audio and embodied cues are intentionally *redundant yet complementary*: speech conveys intent and timing and embodied behaviors (gaze, pointing, proxemics) ground the dialogue in the physical scene. This event-driven policy, tied to navigation and perception events, is designed to reduce working-memory demands and cognitive workload, maintain joint attention, and support successful task completion while preserving user autonomy—an essential factor for effective robotic interaction and assistance for people with dementia.

3.5 Gaze Sensing and Interaction Logging

A lightweight, glasses-style eye tracker ETVision was worn by participants throughout the medication-taking task to unobtrusively capture gaze direction and fixation duration at 180 Hz. This allowed us to observe moments when participants looked at key objects (such as the pill bottle or robot interface) but did not act immediately, which may indicate confusion or difficulty [37]. Each participant completed a brief calibration procedure at the start of the session to ensure accurate gaze mapping. The resulting gaze data were time-stamped and synchronized with the robot’s verbal and gestural outputs to explore the feasibility of detecting potential “confusion events.” While the present system does not yet act on gaze data in real time, data from this work will inform our planned gaze-triggered prompting feature.

4 Formative Usability Study with Healthy Adults

Before recruiting older adults and people living with dementia (PLWDs), we conducted a formative usability study with healthy adults to de-risk the protocol and refine system components. This preliminary phase was motivated by practical and ethical considerations: to surface usability issues prior to testing with vulnerable participants, reduce the likelihood of confusion during task execution, and identify failure modes in robot-assisted medication management that could adversely affect user experience or system performance. We treat this phase as formative rather than confirmatory; its purpose was feasibility and safety tuning, not generalization to the target population. We acknowledge the significant differences between this initial group and PLWDs, and we do not assume that findings fully extrapolate to the target population.

This study was approved by the Institutional Review Board (IRB #####). Inclusion criteria were: (1) age ≥ 18 years and (2) English-speaking. We recruited $N = 12$ adults (ages 18–64) from the university. Six participants reported familiarity with dementia caregiving, and eight reported familiarity with similar technologies prior to participation. All sessions were conducted in a university laboratory setting. Insights from this phase informed subsequent refinements to the protocol and system components prior to recruiting older adults and PLWDs.

4.1 Experimental Procedure and Task Design

Prior to any study activities, participants reviewed and signed an informed consent form authorizing recording and the use of collected study data for academic research. Participants were reminded that participation was voluntary and that they could withdraw at any time if they felt discomfort. After consent, a pre-study survey captured demographics (age, gender, education), background experience with technologies, and prior knowledge or experience with dementia caregiving. The experimenter then assisted each participant in wearing the ETVision gaze tracker and provided instructions describing the session flow. We employed a within-session,

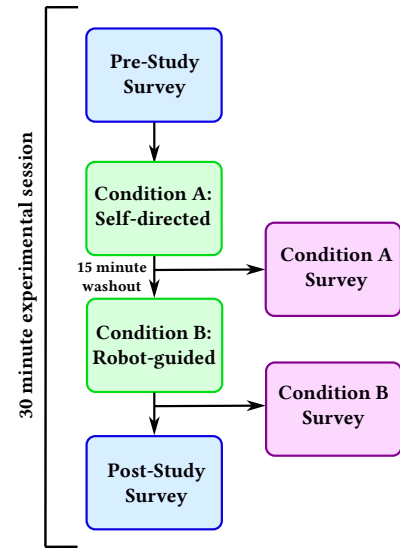


Figure 4: Experimental procedure

two-condition design for a simulated medication-taking scenario (see Figure 4), separated by a 15-minute washout interval, during which the environment was altered. Before each condition, we randomized the pill bottle’s location within a cluttered layout and reconfigured surrounding objects. This was intended to emulate the situational unfamiliarity that PLWDs often experience—even in otherwise familiar spaces—due to memory impairments. The two task conditions were: (A) a self-directed baseline with optional, on-request verbal guidance; and (B) a robot-guided condition using our full navigation–perception–interaction pipeline. In both conditions, the goal was to locate a pill bottle prepared for the study (emptied) and open the cap; opening the cap marked task completion. All sessions were conducted in a university laboratory.

Condition A: Baseline (self-directed). Participants navigated the lab and searched for the pill bottle independently. The robot remained present but passive, providing only brief verbal guidance on request (e.g., possible locations) and occasional encouragement. Guidance could be requested by pressing the *Start Recording* button on the robot’s tablet to speak to the system. Upon completing the task, participants filled out a brief session questionnaire.

Condition B: Robot-guided. (Figure 5). Participants then repeated the activity with robot guidance. The tablet displayed two controls: *Start Navigation* and *Start Recording*. As demonstrated in Figure 5, when *Start Navigation* was pressed, the robot escorted the participant toward a predefined region of interest (ROI) in the lab. After the bottle was detected, the robot oriented its base and head and executed a short pointing gesture to establish joint attention, then prompted the participant to press *Start Recording* to begin the interaction phase. Once the participant indicated that the medication task was complete (i.e., the cap was opened), the robot acknowledged completion and terminated the routine.

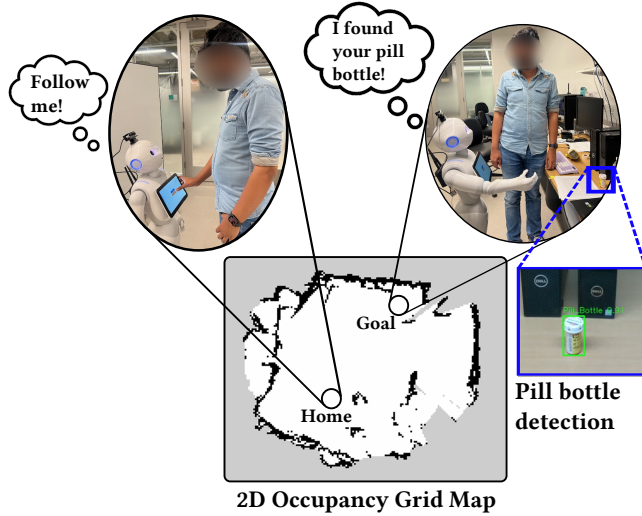


Figure 5: The 2D occupancy grid map represents the lab used for the tests. Image on top left shows the interaction between the participant and the robot. Image on top right shows the interaction between the participant and the robot after it guided and detected the pill bottle.

4.2 Evaluation Metrics

We assessed the system through three complementary approaches to evaluate workload, efficiency, and overall user experience.

Workload Assessment (NASA-TLX). Participants completed the NASA Task Load Index (NASA-TLX) after each condition to assess perceived workload across six dimensions: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. Items were rated on a 1–10 scale. As pairwise weights were not collected, we computed Raw TLX as the unweighted mean of the six dimensions and linearly transformed scores to a 0–100 scale

for interpretability:

$$TLX_{0-100} = \frac{TLX_{raw} - 1}{10 - 1} \times 100.$$

This approach enabled within-participant comparison across assistance levels to determine how cognitive load varied with our adaptive framework.

Behavioral Efficiency (Gaze-Tracking). Eye-tracking data provided objective measures of task performance: (1) time-to-locate-duration from room entry to first verified fixation on the correct pill bottle-and (2) robot-interaction rounds-number of conversational exchanges between participant and robot. These metrics captured both task efficiency and the extent of human-robot engagement.

Perceived Usability (Adapted SUS/QUIS). A post-study questionnaire comprising nine items adapted from the System Usability Scale (SUS) [28] and Questionnaire for User Interaction Satisfaction (QUIS) [4] assessed perceived ease of use, effectiveness, and satisfaction. As this was a formative pilot, we used an adapted subset rather than complete instruments. Items were rated on a five-point Likert scale (1 = strongly disagree, 5 = strongly agree), with items Q2 and Q4 reverse-coded. We computed an adapted usability composite by averaging item scores at the participant level and transforming to a 0–100 scale for consistency with other metrics. We report means with 95% confidence intervals, medians, interquartile ranges, and Cronbach’s alpha for internal consistency.

Qualitative Feedback. Open-ended questions elicited concerns and suggestions (e.g., “One thing about the robot that needs improvement”; “Any concerns about privacy or safety while using the robot for medication management”) to inform iterative design and refinement.

4.3 Ethics, Privacy, and Data Management

All participants provided written informed consent and could withdraw at any time without penalty. Gaze-tracking data were de-identified at the point of collection and stored without direct identifiers. Sessions were video-recorded on a GoPro and held temporarily on the device; audio recordings were deleted immediately after each interaction. After each session, video files were transferred to a secure, access-restricted hard drive available only to IRB-approved study personnel. All identifiable information was handled confidentially in accordance with the consent form and institutional policy.

5 Results and Discussion

Our formative evaluation with healthy adults ($N = 12$, 9 aged 20–30 and 3 aged 30–60) examined whether the adaptive multimodal framework successfully balanced assistance provision with cognitive demand, and how users perceived the system’s usability. Half of the participants (6/12) reported familiarity with people living with dementia or direct dementia caregiving experience. Figure 6 displays individual participant data for workload, task completion time, and robot interaction frequency across both conditions (Condition A: verbal only; Condition B: verbal+gesture+navigation). Results are organized around three key findings that inform future deployment with PLWDs.

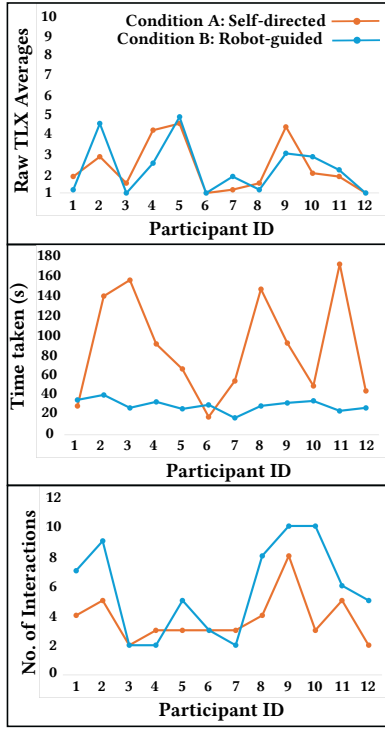


Figure 6: Top plot indicates the mean Raw TLX scores for each participant for each condition. Middle plot shows the time taken by each participant to find the pill bottle for each condition. Bottom plot shows the number of interactions each participant had with the robot for each condition.

5.1 Adaptive Robotic Assistance Reduces Workload Without Overwhelming Users

Raw NASA-TLX scores were uniformly low across conditions (Table 1), with mean normalized values of 20.24/100 for Condition A (verbal only) and 13.89/100 for Condition B (verbal + gesture + navigation). On this 0–100 scale, lower scores indicate lower perceived workload, suggesting a lighter cognitive burden in the embodied condition. While the richer multimodal condition showed numerically lower workload, the small difference and restricted variance (both <25/100) suggest floor effects in this controlled laboratory setting with healthy participants performing relatively simple tasks.

Table 1: Task-level summary: Raw NASA-TLX (0–100), time-to-locate (s), and robot-interaction rounds.

Condition	TLX (M) (0–100)	Time-to-locate (s)	Robot-interaction rounds
Condition A (verbal only)	20.24	91.8	3
Condition B (verbal+gesture+nav)	13.89	29.7	5.5

The low workload across conditions is encouraging: even the most comprehensive assistance level (Condition B) did not overwhelm users cognitively. This suggests that multimodal guidance—combining verbal instructions, gestural cues, and physical navigation—can be delivered without imposing excessive cognitive burden. However, the floor effect limits our ability to detect meaningful differences. Future studies with PLWDs performing more challenging medication regimens in naturalistic home environments will provide a more sensitive test of workload differences across assistance levels.

5.2 Multimodal Robotic Guidance Enhances Task Efficiency and Engagement

Gaze-tracking data revealed substantial practical benefits of embodied assistance (Figure 7). Participants located target medications three times faster with full multimodal support (29.7s) compared to verbal-only reminders (91.8s). Correspondingly, robot-interaction rounds increased from 3 (Condition A) to 5.5 (Condition B), indicating that participants engaged more extensively with the robot when richer guidance was available.

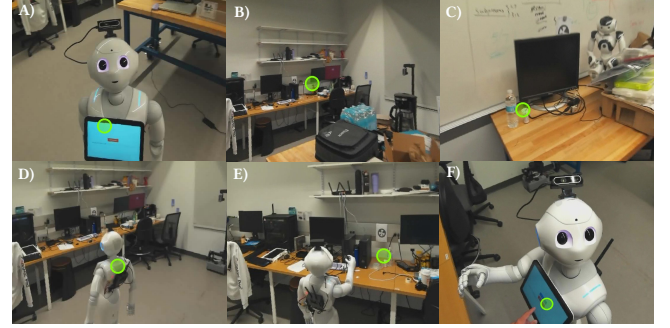


Figure 7: Top row indicates the data collected during Condition A. Bottom row indicates the data collected during Condition B. A) Participant interacts with the robot. B) Participant looks around the room to search for the pill bottle. C) Participant finds the pill bottle. D) Participant is following the robot to pill bottle location. E) Robot gestures towards the pill bottle. F) Participant interacts with robot for further guidance.

These behavioral measures demonstrate that adaptive escalation from minimal to comprehensive assistance produces tangible efficiency gains. For PLWDs who experience spatial disorientation or difficulty executing multi-step tasks [30], faster medication location and more supported interactions, along with step-by-step guidance, could meaningfully impact adherence and reduce caregiver burden. The increased engagement in the multimodal condition aligns with our design goal of providing "just-enough" assistance: users naturally sought more interaction when they received navigational and gestural support, suggesting the system successfully scaffolded task completion rather than replacing user agency.

5.3 High Perceived Usability with Interface Refinement Needs

The adapted usability composite yielded $M=83.10$ (95% CI [70.39, 95.82]), median=88.89, $IQR=6.25$, with high internal consistency ($\alpha = 0.954$). While these scores cannot be directly compared to standard SUS/QUIS benchmarks due to our adapted instrument, the high ratings and tight clustering suggest strong perceived usability. Figure 8 shows item-level distributions, revealing consistently positive responses across all nine questions, with most items showing median scores at or above 4 ("somewhat agree") on the 5-point scale. The reversed items (Q2 and Q4) showed appropriately low scores, indicating participants did not perceive the system as complex or cumbersome.

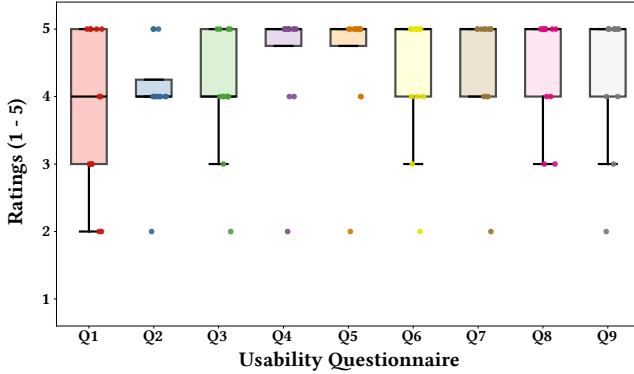


Figure 8: Adapted usability questionnaire: item distributions and composite (0–100).

Qualitative feedback was broadly positive, with participants and stakeholders noting the robot’s potential to support PLWDs during their daily routine activities, promoting their quality of life and independence. However, one consistent concern emerged: confusion around the tablet’s Start/Stop Recording buttons used to initiate interaction. This interface element requires redesign to ensure intuitive operation, particularly for users with cognitive impairments.

The strong usability ratings provide preliminary validation that our adaptive framework is comprehensible and acceptable to users. The high internal consistency ($\alpha \approx 0.95$) indicates our adapted questionnaire reliably captured a coherent usability construct. However, the identified confusion with interaction initiation highlights a critical design consideration for PLWD deployment: interface elements must be unambiguous and cognitively minimal. We plan to replace the current button system with more intuitive triggers (e.g., proximity-based activation or voice-initiated commands) in the next iteration.

5.4 Limitations and Future Work

This pilot ($N = 12$) with healthy adults in a single-room laboratory offers limited statistical power and generalizability to older adults and people with mild cognitive impairment; the simplified setting likely introduced floor effects. Participants reported slow response times and delays in object detection/interaction, but because the

perception–planning–actuation pipeline was not instrumented end-to-end, latency could not be attributed to specific components (e.g., perception, path planning, UI feedback). Finally, workload and usability were assessed using Raw NASA–TLX (unweighted) and an adapted SUS/QUIS subset, so composites should be interpreted as internal measures rather than canonical benchmarks.

We will scale to an adequately powered study that includes older adults and people with mild cognitive impairment and extends evaluation beyond a single-room lab to multi-room home/clinic layouts with occlusions and dynamic obstacles. System priorities include end-to-end latency instrumentation; strengthened motion planning (dynamic obstacle avoidance, narrow-corridor navigation, human-aware pathing, and semantic/topological maps with rapid re-planning under perception dropouts) [5]; privacy-preserving visual monitoring for task state (on-device inference, event-level logs in place of raw video, configurable retention, and explicit consent/UI indicators); and an expanded tablet UI (stepwise cards with large icons, progress/status and error-transparency, multimodal redundancy, and accessible design). We will report effect sizes with confidence intervals and conduct an a priori power analysis to guide sampling.

6 Conclusion

We presented an adaptive, multimodal robotic framework operationalizing occupational therapy’s *assist-as-needed* principle for medication management on a social robot platform. The system links perception to action and dialogue via a unified pipeline: YOLO v11-based detection with depth-guided refinement to recover a 3-D target; joint attention via head alignment and a brief pointing gesture; push-to-talk speech transcription (Whisper); LLM-driven stepwise guidance (GPT-4o); and an event-driven policy escalating from reminders to full multimodal support only as needed—preserving autonomy while maintaining task progress. In a formative lab study with healthy adults and dementia-care stakeholders, participants reported high perceived usability on our adapted instrument ($M \approx 83/100$; $\alpha \approx 0.95$); gaze-derived timing suggested faster target acquisition and greater engagement under richer embodied guidance, while workload remained low across conditions. Open-ended feedback highlighted UI confusion around Start/Stop Recording and perceived latency. These results, while preliminary, indicate feasibility and promise for graded, embodied assistance in medication routines. Limitations include the small sample, controlled setting, and use of adapted measures; accordingly, we plan powered studies with older adults and PLWDs in naturalistic home environments, end-to-end latency instrumentation, strengthened navigation/planning, privacy-preserving visual monitoring for task state, and an expanded, accessible tablet UI. Taken together, this work establishes a theoretically grounded, empirically informed foundation for person-centered, autonomy-preserving robotic assistance in dementia care. By translating occupational therapy principles into adaptive HRI design, our framework advances toward robotic systems that complement rather than replace human caregiving—ultimately empowering independence and dignity for people living with dementia.

References

- [1] Christian K Andersen, Kim U Wittrup-Jensen, Anette Lolk, Kjeld Andersen, and Per Kragh-Sørensen. 2004. Ability to perform activities of daily living is the main factor affecting quality of life in patients with dementia. *Health and quality of life outcomes* 2, 1 (2004), 52.
- [2] Sönke Arlt, Reinhard Lindner, Alexander Rösler, and Wolfgang von Renteln-Kruse. 2008. Adherence to medication in patients with dementia: predictors and strategies for improvement. *Drugs & aging* 25, 12 (2008), 1033–1047.
- [3] SDTMJ Bolarinwa, I Eimontaite, and P Caleb-Solly. 2019. The use of different feedback modalities and verbal collaboration in tele-robotic assistance| IEEE Conference Publication| IEEE Xplore. (2019).
- [4] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.
- [5] Jeel Chatrola, Abhiroop Ajith, Kevin Leahy, and Constantinos Chamzas. 2025. Multi-layer Motion Planning with Kinodynamic and Spatio-Temporal Constraints. In *Proceedings of the 28th ACM International Conference on Hybrid Systems: Computation and Control*. 1–7.
- [6] Dagoberto Cruz-Sandoval, Arturo Morales-Tellez, Eduardo Benitez Sandoval, and Jesus Favela. 2020. A social robot as therapy facilitator in interventions to deal with dementia-related behavioral symptoms. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*. 161–169.
- [7] Alzheimer’s Disease and Dementia. 2024. Alzheimer’s disease facts and figures. *Alzheimers Dement.* 20 (2024), 5.
- [8] Administration for Community Living. 2021. 2020 profile of older Americans. *US department of health and human services* (2021).
- [9] Dieter Fox. 2001. KLD-Sampling: Adaptive Particle Filters. In *Advances in Neural Information Processing Systems 14 (NeurIPS 2001)*. https://papers.nips.cc/paper_files/paper/2001/hash/c5b2cebf15b205503560c4e8e6d1ea78-Abstract.html
- [10] Dieter Fox, Wolfram Burgard, and Sebastian Thrun. 1997. The Dynamic Window Approach to Collision Avoidance. *IEEE Robotics & Automation Magazine* 4, 1 (1997), 23–33.
- [11] Moojan Ghafurian, Jesse Hoey, and Kerstin Dautenhahn. 2021. Social robots for the care of persons with dementia: a systematic review. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 4 (2021), 1–31.
- [12] Birgit Graf, Matthias Hans, and Rolf D Schraft. 2004. Care-O-bot II—Development of a next generation robotic home assistant. *Autonomous robots* 16, 2 (2004), 193–205.
- [13] Giorgio Grisetti, Cyrill Stachniss, and Wolfram Burgard. 2007. Improved Techniques for Grid Mapping with Rao-Blackwellized Particle Filters. *IEEE Transactions on Robotics* 23, 1 (2007), 34–46.
- [14] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.
- [15] Afeez Abiola Hazzan, Jason Dauenhauer, Patti Follansbee, Joyce O Hazzan, Katy Allen, and Ibukun Omobepade. 2022. Family caregiver quality of life and the care provided to older people living with dementia: qualitative analyses of caregiver interviews. *BMC geriatrics* 22, 1 (2022), 86.
- [16] Julian Hirt, Nicola Ballhausen, Alexandra Hering, Matthias Kliegel, Thomas Beer, and Gabriele Meyer. 2021. Social robot interventions for people with dementia: a systematic review on effects and quality of reporting. *Journal of Alzheimer’s Disease* 79, 2 (2021), 773–792.
- [17] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).
- [18] Glenn Jocher and Jing Qiu. 2024. *Ultralytics YOLO11*. <https://github.com/ultralytics/ultralytics>
- [19] Nina Jøranson, Ingeborg Pedersen, Anne Marie Mork Rokstad, and Camilla Ihlebæk. 2016. Change in quality of life in older people with dementia participating in Paro-activity: A cluster-randomized controlled trial. *Journal of advanced nursing* 72, 12 (2016), 3020–3033.
- [20] Chen Ke, Vivian Wei-qun Lou, Kelvin Cheng-kian Tan, Man Yi Wai, and Lai Lok Chan. 2020. Changes in technology acceptance among older people with dementia: the role of social robot engagement. *International journal of medical informatics* 141 (2020), 104241.
- [21] Ioannis Kostavelis, Dimitrios Giakoumis, Georgia Peleka, Andreas Kargakos, Evangelos Skartados, Manolis Vasileiadis, and Dimitrios Tzovaras. 2018. RAMCIP robot: A personal robotic assistant; demonstration of a complete framework. In *Proceedings of the European conference on computer vision (ECCV) workshops*. 0–0.
- [22] Yo-Jen Liao, Ying-Ling Jao, Marie Boltz, Olayemi Timothy Adekeye, Diane Berish, Fengpei Yuan, and Xiaopeng Zhao. 2023. Use of a humanoid robot in supporting dementia care: a qualitative analysis. *SAGE open nursing* 9 (2023), 23779608231179528.
- [23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*.
- [24] Guy M McKhann, David S Knopman, Howard Chertkow, Bradley T Hyman, Clifford R Jack Jr, Claudia H Kawas, William E Klunk, Walter J Koroshetz, Jennifer J Manly, Richard Mayeux, et al. 2011. The diagnosis of dementia due to Alzheimer’s disease: recommendations from the National Institute on Aging-Alzheimer’s Association workgroups on diagnostic guidelines for Alzheimer’s disease. *Alzheimer’s & dementia* 7, 3 (2011), 263–269.
- [25] Christophe Mollaret, Alhayat Ali Mekonnen, Frédéric Lerasle, Isabelle Ferrané, Julien Pinquier, Blandine Boudet, and Pierre Rumeau. 2016. A multi-modal perception based assistive robotic system for the elderly. *Computer Vision and Image Understanding* 149 (2016), 78–97.
- [26] Tyler Morris, Hiroko Dodge, Sylvia Cerel-Suhl, and Xiaopeng Zhao. 2022. Social Robots to support assisted living for persons with Alzheimer’s and related dementias. In *International Conference on Social Robotics*. Springer, 228–237.
- [27] Emilyann Nault, Lynne Baillie, and Frank Broz. 2024. Socially assistive robots and sensory feedback for engaging older adults in cognitive activities. *ACM Transactions on Human-Robot Interaction* 14, 1 (2024), 1–26.
- [28] Kent L Norman, Ben Shneiderman, B Harper, and L Slaughter. 1998. Questionnaire for user interaction satisfaction. *University of Maryland (Norman, 1989)* *Disponível em* 70 (1998).
- [29] Joelle Pineau, Michael Montemerlo, Martha Pollack, Nicholas Roy, and Sebastian Thrun. 2003. Towards robotic assistants in nursing homes: Challenges and results. *Robotics and autonomous systems* 42, 3-4 (2003), 271–281.
- [30] Vinodh Puthusserypady, Laura Emrich-Mills, Emma Lowry, Mital Patel, and Michael Hornberger. 2020. Spatial Disorientation in Alzheimer’s Disease: The Missing Path From Virtual Reality to Real World. *Frontiers in Aging Neuroscience* 12 (2020), 550514. doi:10.3389/fnagi.2020.550514
- [31] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Berger, Rob Wheeler, and Andrew Y. Ng. 2009. ROS: an Open-Source Robot Operating System. In *ICRA Workshop on Open Source Software*. 1–6.
- [32] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356 [eess.AS] <https://arxiv.org/abs/2212.04356>
- [33] Timothy A Salthouse. 2009. When does age-related cognitive decline begin? *Neurobiology of aging* 30, 4 (2009), 507–514.
- [34] Francesca Trojsi, Foteini Christidi, Raffaella Migliaccio, Hernando Santamaria-Garcia, and Gabriella Santangelo. 2018. behavioral and cognitive changes in neurodegenerative diseases and brain injury. *behavioral neurology* 2018 (2018), 4935915.
- [35] Rosalie H Wang, Aishwarya Sudhama, Momotaz Begum, Rajibul Huq, and Alex Mihailidis. 2017. Robots to assist daily activities: views of older adults with Alzheimer’s disease and their caregivers. *International psychogeriatrics* 29, 1 (2017), 67–79.
- [36] Bo Xie, Cui Tao, Juan Li, Robin C Hilsabeck, Alyssa Aguirre, et al. 2020. Artificial intelligence for caregivers of persons with Alzheimer’s disease and related dementias: systematic literature review. *JMIR medical informatics* 8, 8 (2020), e18189.
- [37] Fengpei Yuan, Marie Boltz, Ying-Ling Jao, Arowyn Casenhiser, Aidan Siddiqi, Robert Bray, Joshua Duzan, Monica Crane, and Xiaopeng Zhao. 2022. Assessment of a humanoid partner for older adults and persons with dementia during home-based activities. In *International Conference on Social Robotics*. Springer, 587–597.
- [38] Fengpei Yuan, Robert Bray, Michael Oliver, Joshua Duzan, Monica Crane, and Xiaopeng Zhao. 2024. A social robot-facilitated performance assessment of self-care skills for people with alzheimer’s: A preliminary study. *International Journal of Social Robotics* 16, 9 (2024), 2065–2078.
- [39] Wenzheng Zhao, Kruthika Gangaraju, and Fengpei Yuan. 2025. Multimodal perception-driven decision-making for human-robot interaction: a survey. *Frontiers in Robotics and AI* 12 (2025), 1604472.

A Survey Instruments

A.1 NASA-TLX (condition-based workload)

This appendix reproduces the exact NASA-TLX items used after each task to enable within-participant comparison across assistance types. Items were rated on a 1–10 scale (1 = low, 10 = high). Because pairwise weights were not collected, we computed *Raw TLX* as the unweighted mean of the six items [14] and mapped scores to a 0–100 display scale:

$$TLX_{0-100} = \frac{TLX_{raw} - 1}{10 - 1} \times 100. \quad (1)$$

Table 2: NASA-TLX item wordings and dimensions. Administered after each condition.

Item wording	Dimension
How mentally demanding was the task?	Mental Load
How physically demanding was the task?	Physical Load
How hurried or rushed was the pace of the task?	Temporal Load
How successful were you in accomplishing what you were asked to do?	Performance
How hard did you have to work to accomplish your level of performance?	Effort
How insecure, discouraged, irritated, stressed, and annoyed did you feel?	Frustration

A.2 Adapted usability questionnaire (post-study)

This appendix lists the nine post-study items adapted from SUS and QUIS. Items were answered on a five-point Likert scale (*strongly disagree* to *strongly agree*). Following standard keying, Q2 and Q4 were reverse-coded prior to aggregation. We report an *adapted usability composite* formed as the participant-level mean of all items, mapped to 0–100 for display:

$$U_{0-100} = \frac{U_{raw} - 1}{5 - 1} \times 100, \quad (2)$$

where U_{raw} is the 1–5 mean. Because this is an adapted subset, values are presented as an internal composite and are not compared to canonical SUS/QUIS norms [4, 28].

Table 3: Adapted SUS/QUIS items and mapped constructs (administered once post-study).

ID	Item wording	Construct
Q1	I think that I would like to use the robot frequently.	Usability
Q2	I found the assistance with the robot unnecessarily complex.	Usability (reverse-coded)
Q3	I thought the robot was easy to use.	Perceived ease of use
Q4	I think that I would need the support of a technical person to be able to use the robot.	Perceived ease of use (reverse-coded)
Q5	I would imagine that most people would learn to use the robot very quickly.	Perceived ease of use
Q6	I felt very confident using the robot.	Anxiety
Q7	It is easy to understand the robot's instructions and know what to do.	Information Quality
Q8	I would follow the advice the robot gives me.	Trust
Q9	Overall, I am satisfied with the robot.	Overall Satisfaction

Note: Reliability and descriptive statistics for these adapted items are reported in the main text; item-level distributions are shown in Fig. 8.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009