Robotics: Science and Systems 2024
Delft, Netherlands, July 15-July 19, 2024

Developing Design Guidelines for Older Adults with Robot Learning from Demonstration

Erin Hedlund-Botti, Lakshmi Seelam, Chuxuan Yang, Nathaniel Belles, Zulfiqar Zaidi, and Matthew Gombolay
Georgia Institute of Technology
Atlanta, Georgia, USA
{erin.botti, lseelam3, soyang, nathaniel.belles, zzaidi8}@gatech.edu, matthew.gombolay@cc.gatech.edu

***Abstract*—Assistive in-home robots have the potential to enable older adults to age in place by offloading mentally or physically demanding tasks to a robot. However, one challenge for in-home robots is that each individual will have differing needs, preferences, and home environments, which can all change over time. Learning from Demonstration (LfD) is one solution to enable non-expert users to communicate their differing and changing preferences to a robot, but LfD has not been evaluated with a population of older adults. In a human-subjects experiment where participants teach a robot via LfD, we characterize disparities between older and younger adult participants in terms of robot performance, usability, and participant perceptions. We find that older adults are significantly more critical of the robot's performance and found the LfD process less usable than younger adults. Based on participant performance and feedback, we present design guidelines that will enable roboticists to increase LfD accessibility across demographics.**

I. Introduction

The world's population is aging and with that comes costly in-home care, a shortage of caregivers, and an overburdening of senior living programs [74, 44]. Research has shown that in-home robots have the potential to enable older adults to age in place by providing assistance with routine daily tasks, such as cleaning [76]. This assistance could also reduce the economic burden and labor shortages in the caregiving industry for society as a whole. However, roboticists need to understand how older adults and their potential caregivers perceive robotic technologies, such as what tasks older adults or caregivers would trust a robot to perform and how they would want to program that robot to perform those assistive tasks.

Ideally, a robot would be fully capable out of the box, but that is infeasible given that everyone's home environments, preferences, and needs are different and evolve with time. This is particularly true for older adults whose physical and cognitive abilities may change over time. Therefore, robots will need to be adaptable and personalizable. To make the end-user experience as simple and intuitive as possible, a robot should be able to respond to people's needs in a way that is accessible to people of all levels of technological literacy. In the context of aging in place, older adults will need to be able to communicate their preferences to robots in a way that does not require them to be expert programmers.

Learning from Demonstration (LfD) is one method for enabling non-expert users to teach robots new skills [57]. LfD enables humans to teach robots tasks by demonstrating the desired behavior rather than by manually programming in a scripting language [9]. Therefore, LfD has the potential to enable older adults and caregivers to personalize the behavior of in-home robots. However, limited prior work has investigated LfD with a real target population and, in particular, with older adults. While Paxton et al. [54] developed the Collaborative System for Task Automation and Recognition (CoSTAR), this interface was designed for manufacturing workers with STEM backgrounds rather than older or younger adults without specific training.

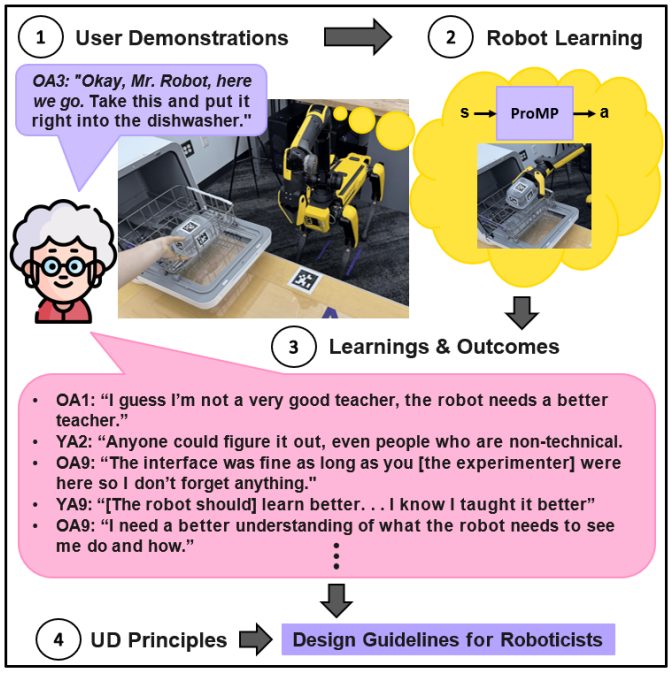

Fig. 1. In our study, older adults (OA) and younger adults (YA) provide demonstrations to a robot. Then the robot learns a policy that maps states, s, to actions, a, via the algorithm, Probabilistic Movement Primitive (ProMP). Participants then view the robot's performance and express feedback on the LfD process. We synthesize the results and contribute design guidelines for how roboticists could improve LfD.

In designing usable systems, prior work has shown the importance of including the target population early into the process to ensure that robotic designs are beneficial and usable by the intended users [18, 73, 50]. Ajaykumar et al. [3] evaluated older adults' opinions on kinesthetically teaching (i.e., physically moving) a robot. In their study, the older adults provided waypoints to the robot and the robot played back the demonstration. In contrast, in our work, the robot

learns in real-time from the users' demonstrations, so we can evaluate participants' perceptions and success with teaching a learning system with stochasticity. Additionally, we investigate a different teaching modality: passive observation, where the robot watches the user do the task, because prior work has found kinesthetic teaching inaccessible for older adults [3, 8]. Saunders et al. [62] created an interface that allowed older adults to provide demonstrations for the robot to learn the user's routine activities; however, their interface was targeted for activity recognition and not LfD. Our work is the first to assess LfD, learning in real-time, with older adults to determine current barriers and potential improvements.

We develop design guidelines to improve the accessibility and usability of LfD by characterizing user interactions with a full-stack LfD robot through the lens of universal design (UD) principles. UD refers to a design practice in which a single design is created to address the needs of heterogeneous users to maximize the number of people who are able and want to use a product [24]. We aim to understand LfD's current barriers for older adults and determine what improvements need to be made for LfD to be widely accessible.

To develop these design guidelines, we first design a novel user interface for our robotic system through participatory, iterative research with our users and guidance from UD principles and usability heuristics [47]. Second, we implement a full-stack LfD robot system and conduct a user study ($n = 32$ participants) where older and younger adults teach the robot via LfD (Fig. 1). We evaluate the robot's performance and participants' perceptions to understand current problems with LfD and characterize disparities in older versus younger adults ($n = 16$ per group). Since most prior work that has evaluated LfD with users recruit populations of convenience, (i.e., young adult university students [73]), we utilize younger adults as a point of comparison. We do not claim that differences between younger and older adults in our experiment will generalize to all older adults but instead characterize key differences to highlight previous perspectives that may have been overlooked. Additionally, our analysis includes covariates, such as experience with technology, to account for potential confounds unrelated to age. Third, we synthesize the results with respect to UD principles to contribute design guidelines for improving accessibility, equity, and performance of LfD for differing populations of end-users: not just one demographic.

II. Related Work

This section discusses UD principles, older adult attitudes about assistive robots, and corresponding LfD techniques.

A. Universal Design (Design for All)

The goal of UD is to design products and systems, such that they are usable by all people, regardless of their age or capabilities. UD includes seven principles: equitable use, flexibility in use, simple and intuitive use, perceptible information, error recovery, low effort, and physical accessibility [24]. UD has been implemented across a variety of fields. In architecture, UD makes an environment safer and healthier, therefore minimizing incidents while allowing people to move freely [20]. Prior work has applied UD practices in human-computer interaction (HCI) applications [69]. For example, Ruzic et al. [61] developed guidelines for web and mobile user interfaces for older adults. HCI researchers have also emphasized the existence of the "evaluation feedback loop" throughout the design and development phases so that users can promptly advocate for their needs [69]. In robotics, UD has been applied to the physical design of spaces to both accommodate robots and humans [43, 58]. UD has been successfully applied in technology and education, resulting in better learning outcomes [66].

Application of UD is especially important for the older adult population, as they are more likely to be challenged by cognitive and memory decline when learning new technologies that could benefit their day to day activities. We build on prior work by investigating LfD with respect to UD principles, especially between older adults and in-home assistive robots.

B. Older Adult Robot Preferences

a) Acceptance of Assistive Robots: Kadylak and Cotten [32] found that while only 24% of older adults were willing to accept in-home robots, their willingness to accept assistive robots was also positively correlated with their limitations in the ability to perform daily tasks. Similarly, Ezer et al. [23] found that for tasks with clear benefits, older adults were more willing to use an in-home robot than younger adults. Additionally, recent surveys show that both older adults and their caregivers are in favor of using care robots [25, 40, 55, 77]. Programming robots could benefit patients with cognitive decline [19]. However, we still need to investigate if older adults are willing and able to teach robots via LfD.

b) Robot Teaching Modality: Older adults' preferences for interacting with the robot are important to consider because there are various ways of teaching a robot using LfD: kinesthetic teaching (i.e., physically manipulating the robot), teleoperation (i.e., remote controlling the robot), and passive observation (i.e., the robot "observing" the participant with a camera) [57]. Fischinger et al. [26] developed the Hobbit robot and employed a multimodal interface for the user to command the robot, and found that older adults preferred voice commands over using a touch screen or gestures.

Similarly, Beer et al. [8] also investigated older adult preferences for controlling a robot, comparing three methods of control: laser pointer, physical manipulation, and devices (e.g., remote control or touch screen). In a user study where older adults observed a robot, older adults were open to a variety of robot control methods, but were wary of physical manipulation. While physical manipulation can let people provide specific instructions, many older adults were concerned about this modality being time-consuming and physically challenging. Furthermore, Ajaykumar et al. [3] found kinesthetic teaching to be inaccessible for older adults. Therefore, in our work, we focus on participants teaching via passive observation.

c) Learning from Demonstration for Assistive Robots: While there are many prior works on LfD methods developed

for assistive tasks, this section focuses on work that includes target end users in assistive domains (e.g., older adults, nurses, and caregivers). Chen and Kemp [14] investigated a kinesthetic teaching interface for nurse assistant robots. Physical correction is common in guiding tasks (e.g., caregivers leading an older adult by the hand) and nurses preferred kinesthetic teaching over teleoperation [14]. Papageorgiou et al. [51] developed an LfD algorithm for human bathing that learns from expert caregivers via passive observation. The learning approach mimics expert demonstrations while taking into account obstacles, for human safety and preference.

Louie and Nejat [41] also implemented an LfD interface that used passive observation. In this study, caregivers taught the robot, Tangy, how to facilitate a game of Bingo for residents at a care facility. Using a similar interface, Saunders et al. [62] conducted a study with a Care-o-Bot3 robot, where participants (three were older adults) taught the robot routines. However, the main focus of the interface was to enable activity recognition by the robot instead of teaching the robot manipulation tasks. Ajaykumar et al. [3] found that older adults preferred familiar and simple interfaces while programming a robot. We build on this work by developing a new LfD interface, with feedback from older adults, and evaluate perceptions of LfD between older and younger adults.

III. Designing Our Study

We utilized a multi-step approach to learn about our target population and inform our main study design. First, we conducted focus groups to identify the potential of home robots for older adults and how older adults would want to teach a robot; these results informed our LfD interface design. Then, we evaluated our interface in a pilot study with older adults. We utilized the lessons learned from the focus groups and pilot study to design an experiment that was pertinent and applicable to our target population.

A. Focus Groups

We conducted two focus groups including individuals with Mild Cognitive Impairment (MCI) and their caregivers, recruited by the Charlie and Harriet Shaffer Cognitive Empowerment Program at Emory University. Focus Group I included 9 participants, four women and five men, and Focus Group II had 13 participants, six women and seven men. Focus Group I examined what tasks older adults want robots to do in their homes. Focus Group II investigated how older adults would want to communicate with a robot. Field notes were taken to analyze participants' attitudes through verbal and non-verbal communication.

We learned from Focus Group I that older adults are excited about ways robotic technologies could assist their daily activities. When the moderator asked participants to brainstorm tasks they would offload to robots, the most common ones mentioned were everyday chores such as cooking, and activities that involve more physical labor and therefore trigger more safety concerns, such as cleaning and doing yard work. Therefore, we use a set of cleaning tasks as relevant domains in our experiments (a finding corroborated by Smarr et al. [68]). The wide range of tasks participants mentioned led to the addition, in our user interface, of a "Task Library," where users can save and name their own tasks. Overall, the results of the focus group analyses suggest positive perceptions toward in-home robots performing assistive tasks.

Focus Group II showed that participants preferred teaching a robot using verbal instructions when this method could sufficiently convey their preferences for the task. However, if tasks required them to convey many details to the robot, participants preferred teaching by showing the task. For example, participants concluded that showing a robot how to clean the sink would help make sure small details are captured, such as how to clean the corners and rims of the sink. Participants preferred physical manipulation the least as they believed it would require too much effort, which is in line with Beer et al. [8] and Ajaykumar et al. [3]. One participant expressed that they would not want to touch the robot. These findings further support the choice to employ passive observation in our study.

B. Interface Design and Pilot Study

Our goal is to create a general purpose LfD interface that takes into consideration user demographics and functional needs when operating a home robot. We opt for a graphic user interface (GUI) as older adults struggle with prescribed language for voice control [35] and the relative familiarity of GUIs may be more accessible [42]. We adapt an existing robot manipulation interface which was developed for in-lab LfD experiments [29]. We designed the GUI by following the UD design principles and Nielson Norman's 10 Usability Heuristics for User Interface Design [46].

For the UD principle of simplicity, we decreased the number of buttons on each page and streamlined user flow by grouping buttons with similar functionality. For example, we separated the interface into two major tabs - record and perform. To teach the robot a new task, the user can go to the record tab and press "Start Record" to begin and "Stop Record" to end. To aid with perceptibility, we increased font and button size to ensure the easy readability of interface elements. Lastly, we added feedback during task completion to reduce users' memory load for progress tracking to incorporate the UD principle for low effort. Tasks are recorded a total of three times before saving, so we included a progress bar at the bottom to keep track of completed recordings. At any point, users can see what the robot is seeing via the "Robot View" window (Fig. 2). When all three demonstrations have been completed, users can press "Save Task" to name the task and save it to the library located underneath the perform tab. In the perform tab, users can drag tasks from the library to the robot's To-Do list. Then, users can press "Play To-Do List" to have the robot execute the list. The perform tab allows users to assemble sub-tasks for the robot to perform longer tasks.

Our interface was employed in a study where university students taught a robot via LfD [45]. We now seek to understand the usability of this interface with older adults. In this work, we

evaluated our interface design using a paper prototype [49] in a pilot study with five participants (aged 55-80, 100% Female). Overall, results were largely positive with participants praising the interface for its minimal, straightforward design. Some critiques included renaming the buttons to terms more familiar to older adults (e.g., queue vs. to-do list) and increasing the level of interaction design. We incorporated this feedback into the final interface design for the main experiment. More details on the pilot study can be found in Seelam et al. [65].

IV. Methodology

We detail our experimental design, including recruitment, domain, research questions, materials, metrics, and procedure.

A. Participants

We recruited two groups of participants to see how perceptions of and success with teaching a robot via LfD differ between age groups. One group is younger adults (aged 20-35) recruited from a university campus. The second group is older adults (aged 60-89) from the local community. Our inclusion criteria were that participants did not have cognitive impairment. We do not currently evaluate our LfD interface with individuals with cognitive impairments, as teaching a new and potentially overwhelming skill might require specific support from caregivers. In future work, we plan to explore techniques to directly incorporate robot instruction from participants who may have diagnosed cognitive impairment.

B. Domain

Based on the results from the focus groups and prior work [68] that support the application of robots for cleaning tasks, we had participants teach a robot using LfD to: 1) place a dish in the dishwasher and 2) use a sponge to wipe the counter. For the dishwasher task, participants taught the robot to place a plastic bowl in the open tabletop dishwasher (Fig. 2). As one motivation for LfD is to personalize robot behavior, participants were instructed to choose their own desired location and orientation of the dish in the dishwasher, as they would when teaching the robot for personal use.

To create a realistic wiping task, a clear piece of acrylic covered the table, on which the experimenter drew a "spill" using a dry-erase marker (Fig. 2). To ensure that the "spills" were consistent across participants, we created stencils of roughly 3-square-inch blob shapes. Participants were instructed to teach the robot to clean the "spill" using a sponge.

C. Research Questions

RQ1: What are the current barriers when teaching a robot with LfD? We investigate if there are any differences between age groups for successful teaching of the robot. We explore how age and additional demographic factors (e.g., personality and attitudes towards technology) impact metrics, such as robot accuracy, participants' ability to use the interface, and types of errors during the teaching process. Then, through qualitative interviews, we query participants about what pain points they experienced when teaching via LfD.

RQ2: How does age group impact participants' perceptions of the robot? We explore whether participants, both older and younger adults, want to teach a robot using LfD to determine if LfD is a viable approach for in-home robots. Through qualitative interviews and quantitative surveys, we evaluate how age group and other demographic factors influence participants' perceptions of LfD.

RQ3: What are potential improvements to the LfD teaching process? In qualitative interviews, we solicit participants' ideas about potential improvements for LfD. We then synthesize the results from RQ1 and RQ2 into a list of design recommendations for LfD systems for older adults.

D. LfD Implementation

We employ the Boston Dynamics Spot robot, which has a 6-DoF arm mounted on a mobile base (Fig. 2). We chose Spot over a stationary tabletop arm because Spot is capable of lifting and manipulating heavy objects while also being able to move around the environment. We wanted participants to envision this robot doing tasks around their homes.

We asked participants for three demonstrations per task, in order to not overload participants. The robot learns via a Probabilistic Movement Primitive (ProMP) algorithm, a specialized form of Movement Primitive tailored for probabilistic modeling of robotic movements [52, 13, 53, 75]. We chose ProMP as they are effective at capturing the variability in demonstrations and can effectively generalize from demonstrated trajectories to novel scenarios [52, 13, 53].

We considered other LfD methods such as behavioral cloning [6], however, behavioral cloning suffers from problems with covariate shift [59] and needs a large amount of data to effectively learn. Furthermore, the study tasks can be described well by movement primitives (e.g., pick-and-place for loading the dishwasher and cyclical/sinusoidal motion for wiping the counter). Methods such as inverse reinforcement learning can learn policies that better account for the human's goals, but these methods require long training times. In contrast, ProMP learns quickly and from few demonstrations. In our study, the median training time for the ProMP was 46 seconds. Lastly, we wanted participants to provide demonstrations, not reward signals or corrective feedback, so we did not employ interactive reinforcement learning or robot-centric LfD methods [36, 59, 38].

ProMPs offer a compact and versatile solution for representing and adapting complex trajectories for dynamic and precise movements. A point on this trajectory, represented as $\mathbf{y}_z \in \mathbb{R}^d$, is formulated as a linear combination of N basis functions, expressed as $\mathbf{y}_z = \Phi_z \boldsymbol{\omega}$. In this expression, $\Phi \in \mathbb{R}^N$ is the matrix of basis functions, $\boldsymbol{\omega} \in \mathbb{R}^{N \times d}$ denotes the adjustable weights, and $z(t) \in [0, 1]$ is the phase variable. ProMPs learn a distribution over the weights, $\boldsymbol{\omega}$, based on multiple demonstrations. This distribution is typically Gaussian, $p(\boldsymbol{\omega}) = \mathcal{N}(\boldsymbol{\omega}; \mu_\omega, \Sigma_\omega)$, with μ_ω and Σ_ω being the mean and covariance matrix determined through maximum likelihood estimation.

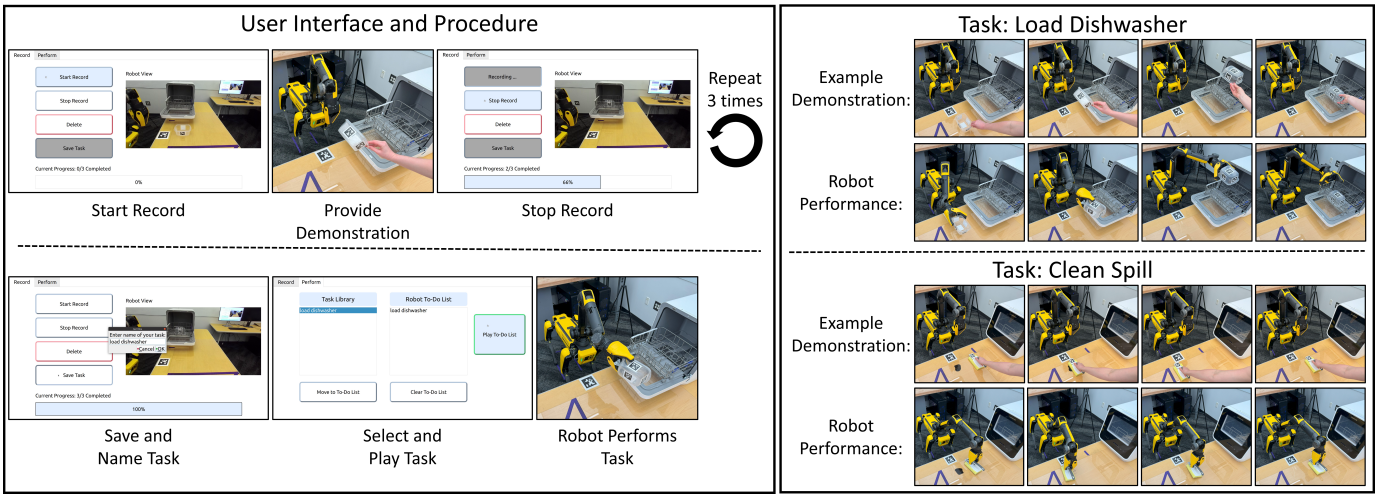

Fig. 2. The left pane displays the procedure for the user interface. The right pane shows the study setup, sample demonstrations, and robot performance.

In our approach, ProMPs are learned through passive observation of demonstrated object trajectories in Cartesian space, aligning with prior findings that suggest users' prefer this teaching mode [8, 3]. Since learning from human video data is an open research question due to the correspondence problem [57], we track the objects using AprilTags [48] instead of tracking the human's hand. To track the AprilTags we employ a Stereolabs ZED2 camera. From each user, for each task, we gather three trajectories, $\{\tau_i\}_{i=1}^{3}$, representing the pose of an object of interest, I, relative to a target frame. Each trajectory, τ_i, is a sequence of poses, $[X_0^I, \ldots, X_{H_i-1}^I]$, where H_i denotes the length of the i-th trajectory. The pose, X_i^I, is defined by the position, $p \in \mathbb{R}^3$, and orientation represented as a quaternion, $q \in \mathbb{R}^4$. An object's positional and orientational trajectory are encoded using object-centric ProMPs. In the generated trajectory, the orientations are normalized to ensure the quaternions' validity. The extension of ProMPs to 3D orientation space is an active area of research [60], and further investigation into this is reserved for future research.

Given that the ProMP is designed to learn only the movement trajectory of the object of interest, I, we predefined the robot's grasping points in relation to the AprilTags positioned on the objects. This approach ensures a consistent and accurate grasp for the execution of the learned trajectories. Once the object is grasped, the robot performs the learned trajectory.

E. Metrics

We now list our metrics with more details in the Appendix. We adhere to the guidelines by Schrum et al. [63] where possible.

1) Pre-Study Metrics: At the beginning of the study, we collect the following metrics.

Demographics: We collect participant's age, gender, race, education, and relationship status through surveys.

Personality: We employ the Mini-IPIP [16] to measure participants' personality traits.

Technology Expertise: We ask participants about their expertise for the following technologies (0=do not use, 1=low, 2=moderate, 3=high): cell phone, tablet, smart home assistant, video calls, social media, computer, GPS (scale range 0-21).

Technology Attitudes: We measure participants' attitudes towards technology [33] (scale range 7-49).

Robot Attitudes: Participants complete the Negative Attitudes towards Robots Scale (NARS) [72].

2) Trial Metrics: For each trial teaching the robot, we collect the following metrics.

Accuracy: We calculate the objective performance of the robot's ability to complete the task. For the wiping the counter task, accuracy is measured by the percentage of the spill cleaned. For loading the dishwasher, we score the accuracy based on if the dish is placed inside the dishwasher and the number/severity of collisions. The dishwasher accuracy was coded by two coders (Cohen's $\kappa = .919$). Accuracy is averaged across tasks and trials.

Robot Similarity: To assess the degree of similarity between the robot's trajectory and the provided demonstrations, we employ dynamic time warping [28] to align the demonstrations with the robot's trajectory. Then, we calculate the similarity between each demonstration and the robot's trajectory [30] and take the average.

Demo Similarity: To evaluate the consistency between one participant's demonstrations, we apply the same method as with Robot Similarity to assess the similarity between pairs of demonstrations, and then take the average.

Perceived Accuracy: After the robot's performance, participants rate the robot (on a scale from 1=Not well at all to 10=Extremely well) on how well the robot completed the task. We measure perceived accuracy to evaluate differences between objective performance and participants' preferences.

Perceived Similarity: After the robot's performance, participants rate the robot (1=Not well at all to 10=Extremely well) on how well the robot matched their demonstration.

Interface Questions: We count the number of times each of the following occurred: a participant made a mistake using the interface, asked the experimenter for a reminder on what to do next, or needed to be prompted with the next step by

the experimenter.

Qualitative Interview: After each trial, we ask participants how the robot did, what they did to teach the robot, and what they wanted the robot to do differently.

3) Post-Study Metrics: After the study, participants evaluated the robot teaching system.

Usability: We employ the System Usability Scale [11].

Reliability and Predictability: Participants complete the Reliability/Competence and Understanding/Predictability subscales from the Trust in Automation scale [37].

Qualitative Interview: We conduct semi-structured interviews to understand participants' overall perceptions of the robot, LfD, and the user interface for usability and potential improvements (see Appendix for list of questions).

F. Procedure

This experiment was approved by our Institutional Review Board (Protocol #H22440). After consenting to having the experiment video recorded, participants complete the pre-surveys to collect demographic information and attitudes towards technology. Participants next watch an instructional video of how to use the interface on an example task: taking out the trash.

Participants then teach the robot two tasks: wiping the counter and loading the dishwasher. Task ordering is randomized and counterbalanced. The instructions inform participants to provide demonstrations slowly and without too much variation. For each task, participants record three demonstrations using the LfD interface. Participants then save and name the task, prompting the robot algorithm to learn. After the robot is done learning, participants use the interface to tell the robot to perform the chosen task. After watching the robot attempt the task, participants evaluate the robot's performance with a post-survey for perceived accuracy and similarity. The experimenter then conducts a short interview asking participants about the robot's performance and teaching process. Participants complete two rounds of teaching the robot and observing the robot's performance for each task. Upon completing the experiment, participants fill out post-surveys about the usability of the teaching system and the reliability/predictability of the robot. Lastly, the experimenter conducts a semi-structured qualitative interview asking participants about their interactions with the system. The study duration is about 1-2 hours and participants are compensated with a $25 Amazon giftcard. Older adults are also compensated for transportation expenses.

V. Results

We report results from 16 older adults (OA), aged 65-80, and 16 younger adults (YA), aged 20-29. Additional participant demographic information is detailed in Table I. The results include both quantitative and qualitative analyses.

For the quantitative analysis, we employ statistical models to evaluate each dependent variable listed in Section IV-E (Trial Metrics and Post-Study Metrics). Before using a parametric model, we test for model assumptions and employ non-parametric tests when appropriate. One aim is to understand how age group impacts participant and robot performance; therefore, a categorical variable for age group is included as the fixed factor in our models. In our analysis of variance (ANOVA), we additionally explore which covariates (e.g., demographics, experience with technology, and attitudes towards robots) impact the dependent variables. We include these covariates to explore potential confounds that could explain the differences between age groups, unrelated to age (e.g., experience with technology). We only include covariates if they lowered the model's AICc score (the Akaike information criterion for small sample sizes). We report significant results with $\alpha < .05$. Details on model composition and assumptions testing are in the Appendix.

	Older Adults	Younger Adults	p-value
Age	70.6 (3.90)	23.9 (2.42)	$\mathbf{p < .001}$
Gender	Female: 56.3% Male: 43.7% Other: 0.0%	Female: 31.3% Male: 68.7% Other: 0.0%	$p = .285$
Race	White: 93.7% Black: 6.3%	White: 18.7% Asian: 81.3%	$\mathbf{p < .001}$
Tech. Expertise	14.4 (3.57)	18.0 (2.24)	$\mathbf{p = .001}$
Tech. Attitudes	30.6 (5.65)	36.9 (5.91)	$\mathbf{p = .005}$

TABLE I
PARTICIPANT DEMOGRAPHICS WITH MEAN (AND STANDARD DEVIATION) OR PERCENTAGE. CONTINUOUS VARIABLES WERE TESTED WITH AN ANOVA OR KRUSKAL-WALLIS TEST, WHILE CATEGORICAL VARIABLES WERE TESTED WITH A χ^2 TEST.

For the qualitative analysis, we report results synthesized using affinity mapping to extract relevant themes from the data. Affinity mapping is a method for organizing large amounts of qualitative data to identify common themes [39]. Each individual point from the qualitative interviews is placed on a separate sticky note and then grouped with others based on similarity. This method provides a visual representation of the qualitative data as a beginning to distilling major themes and motifs. For example, the different tasks which older adults listed as their preferred opportunities for a robot to take over would be grouped according to similarity. This grouping then allows for the researchers to see which tasks provide the most opportunity for home robot intervention. The themes were coded and extracted by two reviewers.

In Section V-A, we present the results categorized based on quantitative metrics and main themes from the qualitative analysis. For each theme, we highlight any similarities or differences between older and younger adults found in our study. Figure 3 showcases the quantitative results, while Table II summarizes important differences between age groups. Then, in Section V-B, we discuss participant reported pain points and potential improvements to the LfD process.

A. Teaching Outcomes and Perceptions of LfD

Positive Feedback: The interviews revealed that most participants found the teaching process to be exciting and fun. Additionally, most participants were comfortable with using the interface and felt that the instructional video was helpful in explaining the functionality of the overall system.

OA1: *"I felt comfortable with [teaching]."*

OA3: *"It was fun. Interesting to see how a robot operates!"*

YA2: *"Anyone could figure it out, even people who are non-technical."*

Value of LfD: Overall, participants expressed that the time investment in teaching a robot through LfD would be worth the value of offloading specific tasks to a robot for an extended period. Other factors, including whether there is on-demand help from experts in the teaching process, or whether the robot can collaborate effectively with all involved parties including caregivers can also influence LfD's perceived value. In addition, while younger adults seemed more eager to offload any household chore they could, older adults overall preferred to only offload tasks they physically could not do.

OA9: *"If I couldn't do them myself, it wouldn't be bad to have a robot."*

YA13: *"The time commitment is worth it, assuming it works."*

Desire to Teach Robot: When asked whether participants wanted to complete the teaching process themselves or if they would want a robot technician to complete the process, the majority of participants shared their desire to teach the robot. Participants expressed that teaching the robot themselves would allow them to have the greatest control and personalization in how the tasks are taught and completed.

An important distinction to make here is that although older adults, in our study, expressed an interest in teaching the robot to support personalization, they saw the benefit of having a robot technician present. Particularly, older adults stated that having a robot technician present to train them on how best to teach the robot would be beneficial.

OA9: *"I prefer to teach myself so I know it's the way I want it to be."*

OA7: *"Maybe a technician for the initial setup, and then teach it myself."*

YA18: *"I would want to do the teaching so it can be personalized."*

Accuracy: We explore both the objective accuracy and participants' perceived accuracy of the robot. ***Nota Bene:*** *Our system was designed to be robust and automatic to give participants a full experience in training a full-stack robot via LfD. However, our system was not perfect as (1) LfD and robot learning are still the domains of research and the laboratory and (2) a system engineered to perfection (if even possible) would have prevented us from characterizing what does and does not work for a variety of end-users. Our results described below indicate that our robot system did achieve the desired level of performance to answer our key research questions.*

We explore which factors impacted the participants' perceived accuracy of the robot. An ANOVA with *perceived* accuracy as the dependent variable found that age group, technology attitudes, relationship status, agreeableness, objective accuracy, and perceived similarity were significant factors. Older adults perceived the robot as less accurate than younger adults ($F(1, 23) = 25.0$, $p < .001$). Participants with more positive attitudes towards technology perceived the robot as less accurate ($F(1, 23) = 14.9$, $p < .001$), which may be due to having higher initial expectations of the robot. Participants scoring higher for the personality trait of agreeableness rated the robot as more accurate ($F(1, 23) = 24.9$, $p < .001$), which could be because participants wanted to be more considerate to the robot or the experimenter. When the robot performed better, participants perceived the robot as more accurate ($F(1, 23) = 20.6$, $p < .001$). Furthermore, when participants thought the robot more closely matched their demonstrations, perceived accuracy of the robot increased ($F(1, 23) = 22.3$, $p < .001$).

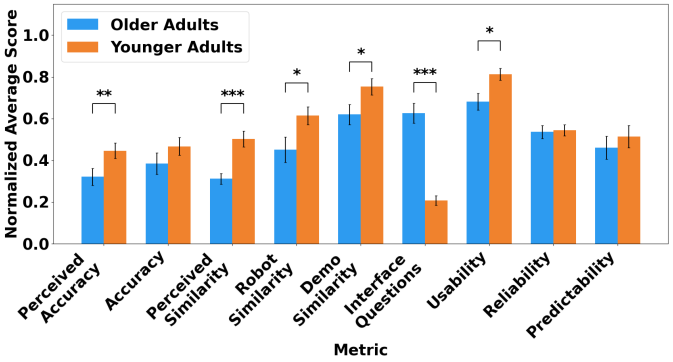

Fig. 3. This figure presents quantitative differences between older and younger adults (*$p < .05$, **$p < .01$, ***$p < .001$). Older adults perceived the robot as less accurate, perceived the robot's trajectories with a lower similarity to their own, and provided more varied demonstrations versus younger adults. Older adults required more assistance with the interface and rated the LfD process as less usable than younger adults.

In addition to quantitatively rating the robot as less accurate, older adults also verbally expressed more criticism towards the robot's performance compared to younger adults.

OA3: *"I don't feel like it did that well because it dropped [the dish] to the dishwasher, that's not what I did"*

YA13: *"It did well, it did what I instructed"* – Note: However, the robot did not fully clean the spill.

While older adults perceived the robot as less accurate than younger adults, an ANOVA with objective accuracy as the dependent variable did not yield any significant differences between age groups. However, participants' demonstration similarity impacted objective accuracy ($F(1, 26) = 6.96$, $p = .014$), meaning that when participants provided more consistent demonstrations, the robot performed better.

Takeaway: Older adults provided more critical feedback of the robot's performance than younger adults.

Similarity: After each trial, participants reported their perceived similarity of how well the robot matched their demonstrations. We also objectively calculate robot similarity – how well the robot matched the person's demonstrations – and demo similarity – measuring the lack of variation in a person's demonstrations. An ANOVA with perceived similarity as the dependent variable found age group and accuracy to be significant factors. Older adults perceived the robot to match their own demonstrations significantly less than younger adults ($F(1, 28) = 25.1$, $p < .001$). The robot's accuracy positively impacted participants' perceived similarity ($F(1, 28) = 11.7$, $p = .002$), meaning participants thought the robot more closely

matched their demonstrations when it performed better.

OA3: *"I made sure to pick it up and put it down gently, which is not what [the robot] did"*

With robot similarity as the dependent variable, an ANOVA reported age group and technology attitudes to be significant. Older adults had significantly lower robot similarity scores compared to younger adults ($F(1, 25) = 7.27$, $p = .012$). Participants with less positive attitudes toward technology were more likely to teach the robot in a way that the robot better matched the person's demonstrations ($F(1, 25) = 4.26$, $p = .049$). It is possible that participants who were more wary of technology were more deliberate with their demonstrations. Additionally, we explored the variation in participants' demonstrations. An ANOVA with demo similarity as the dependent variable found that older adults provided more varied demonstrations than younger adults ($F(1, 28) = 5.39$, $p = .028$). This variation in demonstrations could be why robot similarity scores were lower for older adults.

OA7: *"I picked up the dish, and put it in the dishwasher each time a different way"*

YA15: *"I tried to hold the box the way [the robot's] holding and used more height to put it back"*

Takeaway: Older adults provided more varied demonstrations, impacting robot performance. Participants thought the robot better matched their demonstrations when the robot performed better.

Reliability and Predictability: We found no significant difference between age groups for participants' reliability and predictability ratings of the robot using an ANOVA.

Ease of Use: We evaluate the amount of assistance participants needed and participants' usability ratings of the process.

Interface Questions: A Kruskal-Wallis test with the number of times participants needed help using the interface as the dependent variable found age group to be a significant factor ($\chi^2(1) = 22.8$, $p < .001$). Older adults needed significantly more assistance from the experimenter to correctly use the interface compared to younger adults.

Usability: An ANOVA with usability as the dependent variable found age group, technology expertise, education, and accuracy all to be significant factors. Older adults found the teaching process to be significantly less usable than younger adults ($F(1, 23) = 14.9$, $p < .001$). Mean SUS scores were 68.1 (range 37.5-87.5) for older adults and 81.3 (range 55.0-100.0) for younger adults. Participants who reported more expertise with technology found the system more usable ($F(1, 23) = 9.61$, $p = .005$). Unsurprisingly, when the robot performed more accurately, participants found the system to be more usable ($F(1, 23) = 5.43$, $p = .029$).

OA9: *"The interface was fine as long as you [the experimenter] were here so I don't forget anything."*

YA13: *"The interface was very intuitive to use."*

OA9: *"I need a better understanding of what the robot needs to see me do and how."*

YA9: *"Maybe it could have an instruction manual on how the robot will understand your motions. That insight would help us to teach it or like record that motion properly."*

Takeaway: Accuracy of robot task execution is an important factor on how people perceive usability. Also, for older adults to find the system usable, more effort needs to be put into interfaces, instructions, or tutorials on how to teach the robot.

Privacy Concerns: When asked about being recorded at home by the robot, most participants expressed that as long as they had control over the camera, recording would be acceptable, especially if the cameras are needed for learning.

OA2: *"Cameras are fine [if] I'm aware that they are on."*

YA15: *"Recording is fine if the frame is just...the task."*

Form Factor: As Spot was not developed for in-home use, we wanted to understand how participants felt about the form factor of the robot. As a point of comparison, we showed participants a photo of the Stretch [34], a robot designed for in-home scenarios. While older adults generally preferred for in-home robots to have lifelike features (e.g., Spot's dog-like traits), younger adults preferences were split between anthropomorphic vs. machine-like forms. Most participants in our study considered the large size of the robot a limitation for in-home use; however, participants said this also depends on the robot's ability to complete more of their desired tasks.

YA7: *"The less humanised [the robot is], the better."*

OA3: *"I kinda like how it looks like a person or animal."*

B. Barriers and Improvements

Characterizing Teaching Behavior: The overall robot accuracy was relatively low (46.6% for younger adults and 38.5% for older adults). We analyzed the different teaching behaviors or errors that people made in their demonstrations that affected the accuracy of the robot. We note that while some of these teaching behaviors resulted in robot errors, we highlight these results, not to blame participants, but to communicate challenges that LfD designers should consider.

Speed: Participants were given verbal instructions to move slowly while providing demonstrations. Yet, many exhibited faster movements, causing the robot to detect objects less frequently. For the dishwasher task, this caused the robot to take a more direct path and increased the possibility of collision with the dishwasher. This also resulted in the robot having shorter, less effective wiping motions with the sponge.

AprilTags Not Visible: One-fourth of younger adults and half of older adults covered the AprilTags or moved the object out of frame. This reduced object detections and caused the robot to perform worse. This type of error is not only due to our use of AprilTags; occlusions can also impact the accuracy of computer vision object detectors.

Variation in Demonstrations: Participants were instructed to keep their demonstrations mostly consistent. Older adults provided more varied demonstrations than younger adults ($p < .001$). For ProMP, the performance is best when demonstrations are similar. However, this problem may not generalize to all LfD methods, and could be a positive for other methods.

Correspondence Problem: Participants would turn the dish or the sponge in orientations difficult for the robot to replicate.

Theme	Older Adults	Younger Adults	Difference
Teaching Confidence	"I guess I'm not a very good teacher, the robot needs a better teacher."	"[The robot should] learn better...I taught it better."	When the robot failed, older adults were more likely to blame themselves, and younger adults more often blamed the robot.
Teaching Modality	Many older adults gave the robot additional, verbal commands, and a few older adults used gestures (e.g., pointing at the sponge or holding the sponge close to the robot's camera for "emphasis").	Younger adults tended to only provide demonstrations, without extraneous input.	Some older adults preferred to provide multi-modal demonstrations (e.g., voice), despite the experimenter explaining that there would be no benefit.
Desire to Teach Robot	"Maybe a technician for the initial setup, and then teach it myself"	"I would want to do the teaching so it can be personalized."	Younger adults wanted to teach the robot themselves, while older adults expressed wanting a technician's help to get started.
Value of LfD	"If I couldn't do them myself, it wouldn't be bad to have a robot."	"The time commitment is worth it, assuming it works."	Older adults overall preferred to only offload tasks they physically cannot do.
Robot Performance	"I don't feel like it did that well because it dropped [the dish] to the dishwasher, that's not what I did"	Despite the robot not fully cleaning the spill, a younger adult commented, "It did well, it did what I instructed"	Older adults perceived the robot's accuracy as lower and were verbally more critical.
Interface Usability	"The interface was fine as long as you [the experimenter] were here so I don't forget anything."	"The interface was very intuitive to use."	Older adults required more instructions and assistance in learning how to teach the robot and use the interface.

TABLE II
THIS TABLE HIGHLIGHTS THE PERCEPTIONS OF LFD THAT DIFFERED BETWEEN PARTICIPANT AGE GROUPS FOUND IN OUR EXPERIMENT.

For example, rotating a dish at a certain angle caused the robot's elbow to collide with the dishwasher door. During interviews, participants proposed alternative grasping methods which, the robot would have been incapable of completing.

Extra Gestures: Some participants used additional gestures to convey information to the robot. Yet, as the robot couldn't interpret these gestures, they disrupted its learning. For example, multiple participants moved the sponge close to the robot's camera, pointed at the sponge and said "Spot, this is the sponge." This led the robot to mistakenly lift the sponge instead of cleaning the spill.

Our fault: On five (out of 64) dishwasher trials, the robot completed the task, but collided with the dishwasher on exit because the home arm command had no obstacle avoidance.

Reported Problems: We report the problems and pain points that participants expressed during the teaching process.

Grasping: Participants repeatedly mentioned that the way the robot grasped the object made the task unsuccessful, and they did not like it when the robot failed to grasp the object in the same way they had demonstrated. It is interesting that this aspect stood out to participants since grasping was pre-programmed by the experimenter, due to the correspondence problem remaining an open challenge. For example, Spot has no fingers and may need to grasp differently than a person to successfully complete the task. Though participants wanted Spot to mimic their grasp, the way they often grasped the object would have been impossible for Spot or would have caused Spot to fail the task later.

OA1: *"I think part of the problem is the handle not matching the grabbing tool of the robot."*

YA8: *"Gripping the angle of the sponge must be different."*

Teaching Confidence: When asked to wipe the spill on the table, participants showed the robot this task in a variety of ways. Some individuals wiped in a circular motion whereas others wiped horizontally. Some completed the entire task in one fluid motion whereas others broke the task down into sequential sub-tasks. As there exists a range of possibilities in demonstrations, participants expressed doubt and confusion on how best to display the task for the robot. Additionally, older adults were less confident in their teaching abilities expressing that the robot may have needed a better teacher to learn the task while younger adults placed more blame on the robot for its inability to learn from their demonstrations.

OA1: *"I guess I'm not a very good teacher, the robot needs a better teacher."*

YA9: *"[The robot should] learn better... I taught it better"*

Understanding Robot Failures: Participants voiced their confusion around not understanding why the robot was unable to complete the task. When the robot fails at the intended task, participants are not given any kind of feedback as to why this failure occurs. As a result, most participants took corrective actions based on their own interpretations on the second attempts of teaching a specific task without clear guidance. Overall, participants expressed that they would want to know whether it was specific parts of their demonstration that needed improvement, or if there was an error on the robot's end, to ensure greater success in learning outcomes.

OA3: *"I tried to be gentle with the dish - why did [the robot] do worse?"*

YA6: *"Maybe knowing how the robot is programmed will help me figure out how to best teach it."*

Spatial Awareness: During the dishwasher task, the robot needed to place the dish in the center of the tray while not colliding with other elements in the environment. In many trials, the robot collided with other elements including the dishwasher tray, door, and table. In these instances, participants expressed the need for the robot to have greater mapping and spatial awareness capabilities.

OA2: *"I would want [the robot] to be more aware spatially and complete the actions in distinct steps."*

Goal Oriented Learning: As the robot is programmed to watch the trajectory of a moving object and then replicate

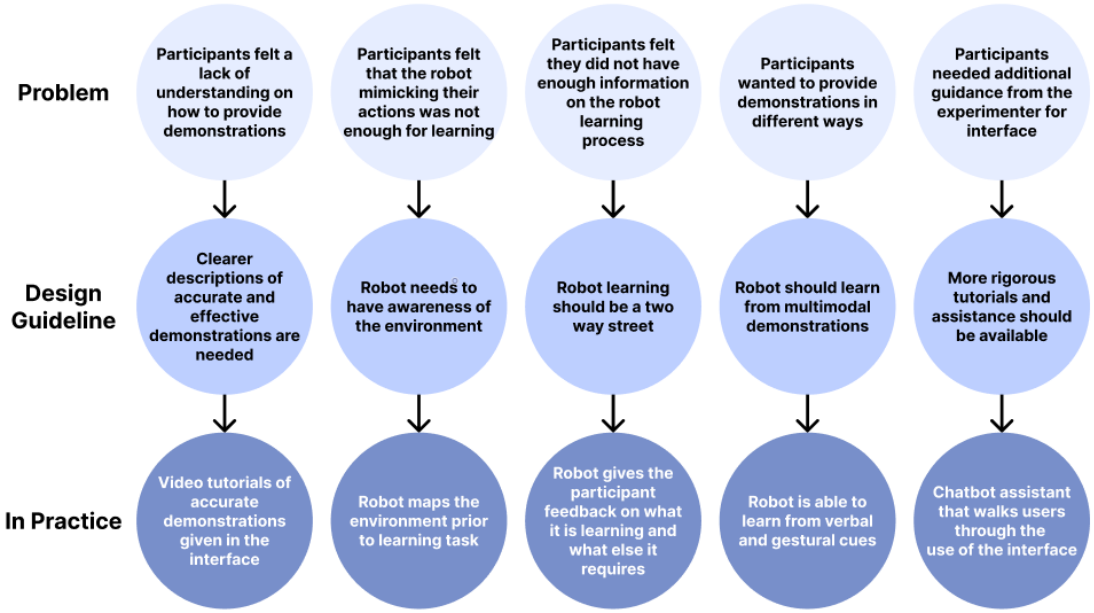

Fig. 4. This figure displays current barriers to teaching with LfD along with design guidelines for potential improvements.

this trajectory, participants found the sole mimicking of their actions frustrating. This feeling resulted from the participants' dissatisfaction with the robot not understanding the goal of their motion (i.e., continuing to wipe until the spill is clean).

YA7: *"If [the robot] can just make sure that the goal was achieved after it mimicked the task that I showed it."*

Potential Improvements: We present feedback from participants on potential improvements to the LfD teaching process. Figure 4 highlights how problems presented in the previous section can be addressed by our design guidelines.

In order to mitigate the confusion on what demonstrations should look like, participants put forth multiple solutions. Some participants expressed a desire for the robot technician to teach them how to complete the demonstrations rather than to teach the robot the task. Other participants shared that receiving more concrete information on how the robot learns and how best to communicate the task would be helpful.

YA10: *"Maybe it could have an instruction manual"*

Throughout the teaching process, participants communicated that they would want feedback on their demonstrations to know what the robot was learning from them. Participants felt that by knowing which part of their demonstration was learned correctly, they could alter their subsequent demonstrations to ensure that the robot was able to complete the entire task.

Participants shared that there is a need for the robot to be aware of more than the participant's motions. These elements include the surrounding environment, obstacles in the way of the task, and the goal of the task. Participants want an avenue to communicate these various elements to the robot in addition to the topography of the task. For example, YA7 voiced that if they were able to communicate the goal of the task, then the robot would be able to ensure that the goal was achieved after the actions were completed.

VI. Discussion

Overall, older and younger adults were both positive about the value of LfD and a willingness to teach a robot, validating LfD as an impactful research direction for use with in-home robots and older adults. Additionally, we found significant differences in how older adults perceived the LfD process compared to younger adults for the participants in this experiment. Older adults were more critical of the robot's performance, less confident while teaching, found the system less usable, and provided more multi-modal input than younger adults. Based on our findings, we present the following design guidelines to consider when developing an LfD system, which we contextualize within UD principles [24] to support the objective of equitable LfD for any age group.

Better Instructions for Demonstrations: Some participants' teaching behaviors (e.g., moving too fast, providing demonstrations the robot could not physically achieve, or adding extraneous gestures in their demonstrations) caused poor robot performance. When the robot failed, some participants expressed confusion about what they did wrong. Teaching people how to be better robot demonstrators is one way to improve robot performance, and there are multiple prior works that investigate ways to teach humans to do LfD [45, 64, 12, 27, 2]. We suggest that designers of LfD systems consider how best to introduce the human to the LfD process and teach humans how to teach robots. Researchers also need to develop principles for interaction design to decide when a robot should be adaptive to human limitations versus when the focus should be on better training for human users [64, 27]. Clearer instructions for how to provide demonstrations will adhere to the UD principle of simplicity: enabling LfD to be more intuitive, independent of one's prior knowledge.

Feedback from the Robot: Participants expressed frustra-

tion at not understanding why the robot failed. More feedback from the robot or explainability techniques could help support participants. Potential areas for improvement include explainable artificial intelligence to communicate reasons for failures [17], more legible motion from the robot [21], and feedback to the person on what the robot understands from their demonstrations. For example, many participants unknowingly covered the AprilTags during demonstrations. Furthermore, the robot being able to show its learned behavior in simulation would also improve safety in an inherently trial-and-error process. Additionally, active learning methods could be useful to prompt the user for specific types of demonstrations [15]. Feedback from the robot will aid in satisfying the UD principles for error recovery and perceptibility, by communicating essential information.

Understanding of Environment and Goal: The simplest version of LfD is behavioral cloning [6], which lacks robustness to changes in the environment and does not take into account the task goal. For example, inverse reinforcement learning [7] techniques attempt to learn reward functions for human goals; however these methods are sample and computationally inefficient. We need more LfD methods that can learn to understand the environment and goals in a reasonable amount of time for participants to be able to teach the robot in real-time [22, 4]. Additionally, methods that enable participants to communicate the importance of specific features, i.e., the sponge should be flat, are necessary [10]. Methods that can better understand the environmental features and task goals will improve adherence to the UD principle for error recovery. Also, developing these methods that can learn in a realistic amount of time for real-world, real-time interactions will assist with the UD principle for low effort by minimizing number of demonstrations that the end-user needs to provide and the amount of time the end-user needs to wait.

Better Tutorial for Interface: While younger adults expressed that the interface was intuitive and easy to use, older adults struggled more. Older adults said that the interface was straightforward once they had some practice, but they needed initial guidance from the experimenter. As such, a tutorial that walked participants through the interface with tips, guidance, and error handling, would be necessary for real-world deployments [1]. Adding a more detailed tutorial will improve LfD with respect to the UD principles of 1) simplicity, by not assuming users have prior knowledge and providing helpful cues to the user, 2) perceptibility, by increasing legibility and adding more modes of communication (e.g., verbal and visual reminders), and 3) error recovery.

Learn from Multi-modal Demonstrations: In addition to demonstrations, participants wanted to talk to the robot and use gestures. For example, participants would point at the object the robot is supposed to manipulate or participants would gesture to the "spill" to indicate that the spill is an important feature. As such, LfD methods should incorporate such multi-modal data [5, 71, 67, 70]. This goal for multi-modal LfD is mirrored by a UD principle: be flexible to user preferences in interaction. By accommodating different types of user input, LfD methods would increase flexibility. Additionally, learning from multi-modal demonstrations could help with physical accessibility. While we only conducted our study with able-bodied participants, allowing multiple ways for participants to provide demonstrations and input could make LfD more accessible for people who can only partially or not fully physically complete the task.

A. Limitations and Future Work

While we conducted our study with multiple age groups, the sample size was small. Due to the small sample size and exploratory nature of the statistical analysis, our results might include some false positives or negatives. In the future, we aim to conduct this study with a larger and more diverse population. Nonetheless, this study represents a significant and novel contribution to robotics owing to the lack of research with target populations of interest [73] and the design guidelines and insights we glean from our mixed-methods investigation.

Another limitation is that there is a novelty effect in our study because many of our older adult participants expressed that they had not interacted with a robot before. Many of the positive comments from participants about teaching the robot included words such as "fun," "interesting," and "engaging." It is possible that teaching the robot may become more cumbersome if participants were teaching it everyday in their homes. Additionally, since the robot's objective accuracy was less than 50%, this can impact perceived metrics such as accuracy and usability [31]. Future work should evaluate if differences in perceptions between older and younger adults hold when the objective accuracy increases.

Comparing the Spot robot to a photo of Stretch was exploratory [56]. Furthermore, our study has a cohort effect; age differences could be due to generational differences. Future work needs to investigate if these results apply to future generations of older adults. There could be additional confounds related to recruitment biases or other differences between populations. Lastly, all of our participants were able-bodied and more research needs to be done for UD of LfD for people with disabilities. Despite these limitations, we hope that this serves as a guide and call to action for more investigation into UD for LfD.

VII. Conclusion

We contribute design guidelines for improving LfD in terms of accuracy and usability, grounded in the application of assistive robotics for older adults. We conducted a user study where participants taught a full-stack LfD robot, and we compared results between older adults – our target population – and younger adults. Older adults found the system less usable and were more critical of their performance as teachers and the robot's performance as a learner compared to younger adults. We propose guidelines based on UD principles for user experience, robot learning, and robot communication to increase accessibility of LfD for all end-users.

ACKNOWLEDGEMENTS

This work was supported by a grant from the National Science Foundation (IIS-2112633), a gift from Konica Minolta, Inc, and the P.E.O. Scholar Award.

REFERENCES

[1] Gopika Ajaykumar, Maia Stiber, and Chien-Ming Huang. Designing user-centric programming aids for kinesthetic teaching of collaborative robots. *Robotics and Autonomous Systems*, 145:103845, 2021. ISSN 0921-8890. doi: https://doi.org/10.1016/j.robot.2021.103845. URL https://www.sciencedirect.com/science/article/pii/S0921889021001305.

[2] Gopika Ajaykumar, Gregory D. Hager, and Chien-Ming Huang. Curricula for teaching end-users to kinesthetically program collaborative robots. *PLOS ONE*, 18 (12):1–20, December 2023. doi: 10.1371/journal.pone.0294786. URL https://doi.org/10.1371/journal.pone.0294786. Publisher: Public Library of Science.

[3] Gopika Ajaykumar, Kaitlynn Taylor Pineda, and Chien-Ming Huang. Older adults' expectations, experiences, and preferences in programming physical robot assistance. *International Journal of Human-Computer Studies*, 180:103127, 2023. ISSN 1071-5819. doi: https://doi.org/10.1016/j.ijhcs.2023.103127. URL https://www.sciencedirect.com/science/article/pii/S1071581923001362.

[4] Baris Akgun and Andrea Thomaz. Simultaneously learning actions and goals from demonstration. *Autonomous Robots*, 40(2):211–227, February 2016. ISSN 1573-7527. doi: 10.1007/s10514-015-9448-x. URL https://doi.org/10.1007/s10514-015-9448-x.

[5] Baris Akgun, Maya Cakmak, Jae Wook Yoo, and Andrea Lockerd Thomaz. Trajectories and keyframes for kinesthetic teaching: a human-robot interaction perspective. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '12, page 391–398, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450310635. doi: 10.1145/2157689.2157815. URL https://doi.org/10.1145/2157689.2157815.

[6] Brenna Argall, Sonia Chernova, Manuela M. Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5):469–483, May 2009. doi: 10.1016/j.robot.2008.10.024. URL https://doi.org/10.1016/j.robot.2008.10.024.

[7] Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2021.103500. URL https://www.sciencedirect.com/science/article/pii/S0004370221000515.

[8] Jenay M. Beer, Akanksha Prakash, Cory-Ann Smarr, Tracy L. Mitzner, Charles C. Kemp, and Wendy A. Rogers. "Commanding Your Robot" Older Adults' Preferences for Methods of Robot Control. *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting. Human Factors and Ergonomics Society. Annual meeting*, 56(1):1263–1267, September 2012. ISSN 1071-1813 2169-5067. doi: 10.1177/1071181312561224.

[9] A. Billard and D. Grollman. Robot learning by demonstration. *Scholarpedia*, 8(12):3824, 2013. doi: 10.4249/scholarpedia.3824. revision #138061.

[10] Andreea Bobu, Marius Wiggert, Claire Tomlin, and Anca D. Dragan. Inducing structure in reward learning by learning features. *The International Journal of Robotics Research*, 41(5):497–518, 2022. doi: 10.1177/02783649221078031. URL https://doi.org/10.1177/02783649221078031. _eprint: https://doi.org/10.1177/02783649221078031.

[11] John Brooke. Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, 189, 11 1995.

[12] Maya Cakmak and Leila Takayama. Teaching People How to Teach Robots: The Effect of Instructional Materials and Dialog Design. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '14, pages 431–438, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 978-1-4503-2658-2. doi: 10.1145/2559636.2559675. URL https://doi.org/10.1145/2559636.2559675. event-place: Bielefeld, Germany.

[13] João Carvalho, Dorothea Koert, Marek Daniv, and Jan Peters. Adapting object-centric probabilistic movement primitives with residual reinforcement learning. In *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, pages 405–412, 2022. doi: 10.1109/Humanoids53995.2022.10000148.

[14] Tiffany L. Chen and Charles C. Kemp. Lead me by the hand: Evaluation of a direct physical interface for nursing assistant robots. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 367–374, 2010. doi: 10.1109/HRI.2010.5453162.

[15] Sonia Chernova and Manuela M. Veloso. Interactive Policy Learning through Confidence-Based Autonomy. *Journal of Artificial Intelligence Research*, 34:1–25, 2009. doi: doi:10.1613/jair.2584.

[16] Andrew J. Cooper, Luke D. Smillie, and Philip J. Corr. A confirmatory factor analysis of the Mini-IPIP five-factor model personality scale. *Personality and Individual Differences*, 48(5):688–691, 2010. ISSN 0191-8869. doi: https://doi.org/10.1016/j.paid.2010.01.004. URL https://www.sciencedirect.com/science/article/pii/S019188691000022X.

[17] Devleena Das, Siddhartha Banerjee, and Sonia Chernova. Explainable ai for robot failures: Generating explanations that improve user assistance in fault recovery. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21, page 351–360, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450382892. doi: 10.1145/3434073.3444657. URL https://doi.org/10.1145/3434073.3444657.

[18] Stevienna de Saille, Eva Kipnis, Stephen Potter, David Cameron, Calum J. R. Webb, Peter Winter, Peter O'Neill, Richard Gold, Kate Halliwell, Lyuba Alboul, Andy J. Bell, Andrew Stratton, and Jon McNamara. Improving Inclusivity in Robotics Design: An Exploration of Methods for Upstream Co-Creation. *Frontiers in Robotics and AI*, 9, 2022. ISSN 2296-9144. doi: 10.3389/frobt.2022.731006. URL https://www.frontiersin.org/articles/10.3389/frobt.2022.731006.

[19] Stavros Demetriadis, Thrasyvoulos Tsiatsos, Theodosios Sapounidis, Magda Tsolaki, and Alexandros Gerontidis. Exploring the potential of programming tasks to benefit patients with mild cognitive impairment. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '16, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450343374. doi: 10.1145/2910674.2935850. URL https://doi.org/10.1145/2910674.2935850.

[20] Halime Demirkan. Housing for the aging population. *European review of aging and physical activity*, 4(1):33–38, 2007.

[21] Anca D. Dragan, Kenton C.T. Lee, and Siddhartha S. Srinivasa. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 301–308, 2013. doi: 10.1109/HRI.2013.6483603.

[22] Cem Eteke, Doğancan Kebüde, and Barış Akgün. Reward learning from very few demonstrations. *IEEE Transactions on Robotics*, 37(3):893–904, 2021. doi: 10.1109/TRO.2020.3038698.

[23] Neta Ezer, Arthur D. Fisk, and Wendy A. Rogers. More than a Servant: Self-Reported Willingness of Younger and Older Adults to having a Robot perform Interactive and Critical Tasks in the Home. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 53(2):136–140, 2009. doi: 10.1177/154193120905300206. URL https://doi.org/10.1177/154193120905300206. _eprint: https://doi.org/10.1177/154193120905300206.

[24] Miranda A Farage, Kenneth W Miller, Funmi Ajayi, and Deborah Hutchins. Design principles to accommodate older adults. *Global journal of health science*, 4(2):2, 2012.

[25] Laura Fiorini, Marleen De Mul, Isabelle Fabbricotti, Raffaele Limosani, Alessandra Vitanza, Grazia D'Onofrio, Michael Tsui, Daniele Sancarlo, Francesco Giuliani, Antonio Greco, et al. Assistive robots to improve the independent living of older persons: results from a needs study. *Disability and Rehabilitation: Assistive Technology*, 16(1):92–102, 2021.

[26] David Fischinger, Peter Einramhof, Konstantinos Papoutsakis, Walter Wohlkinger, Peter Mayer, Paul Panek, Stefan Hofmann, Tobias Koertner, Astrid Weiss, Antonis Argyros, and Markus Vincze. Hobbit, a care robot supporting independent living at home: First prototype and lessons learned. *Robotics and Autonomous Systems*, 75:60–78, 2016. ISSN 0921-8890. doi: https://doi.org/10.1016/j.robot.2014.09.029. URL https://www.sciencedirect.com/science/article/pii/S0921889014002140.

[27] Kanishk Gandhi, Siddharth Karamcheti, Madeline Liao, and Dorsa Sadigh. Eliciting compatible demonstrations for multi-human imitation learning. In Karen Liu, Dana Kulic, and Jeff Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 1981–1991. PMLR, 14–18 Dec 2023. URL https://proceedings.mlr.press/v205/gandhi23a.html.

[28] Toni Giorgino. Computing and visualizing dynamic time warping alignments in r: The dtw package. *Journal of Statistical Software*, 31(7):1–24, 2009. doi: 10.18637/jss.v031.i07. URL https://www.jstatsoft.org/index.php/jss/article/view/v031i07.

[29] Nakul Gopalan, Nina Moorman, Manisha Natarajan, and Matthew Gombolay. Negative Result for Learning from Demonstration: Challenges for End-Users Teaching Robots with Task And Motion Planning Abstractions. In *Proceedings of Robotics: Science and Systems*, New York City, NY, USA, June 2022. doi: 10.15607/RSS.2022.XVIII.028.

[30] Michael Grupp. evo: Python package for the evaluation of odometry and slam. https://github.com/MichaelGrupp/evo, 2017.

[31] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y.C. Chen, Ewart J. De Visser, and Raja Parasuraman. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors*, 53 (5):517–527, 2011. ISSN 00187208. doi: 10.1177/0018720811417254. ISBN: 0018720811417.

[32] Travis Kadylak and Shelia R. Cotten. United States older adults' willingness to use emerging technologies. *Information, Communication & Society*, 23(5):736–750, 2020. doi: 10.1080/1369118X.2020.1713848. URL https://doi.org/10.1080/1369118X.2020.1713848. Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2020.1713848.

[33] Takeo Kanade and Rory A. Cooper. Year 3 annual report and renewal proposal. Technical report, Carnegie Mellon University, 2009. URL http://www.cs.cmu.edu/~cga/qolt/old/QoLT-Yr3-AnnRpt-V1_lite.pdf.

[34] Charles C. Kemp, Aaron Edsinger, Henry M. Clever, and Blaine Matulevich. The design of stretch: A compact, lightweight mobile manipulator for indoor human environments. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 3150–3157, 2022. doi: 10.1109/ICRA46639.2022.9811922.

[35] Sunyoung Kim. Exploring How Older Adults Use a Smart Speaker-Based Voice Assistant in Their First Interactions: Qualitative Study. *JMIR mHealth and uHealth*, 9(1):e20427, January 2021. ISSN 2291-5222. doi: 10.2196/20427. Place: Canada.

[36] Bradley Knox and Peter Stone. TAMER: Training an Agent Manually via Evaluative Reinforcement. In *2008 7th IEEE International Conference on Development and Learning*, pages 292–297, 2008. doi: 10.1109/DEVLRN.2008.4640845.

[37] Moritz Körber. Theoretical considerations and development of a questionnaire to measure trust in automation. March 2018.

[38] Michael Laskey, Caleb Chuck, Jonathan Lee, Jeffrey Mahler, Sanjay Krishnan, Kevin Jamieson, Anca Dragan, and Ken Goldberg. Comparing human-centric and robot-centric sampling for robot deep learning from demonstrations. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 358–365, 2017. doi: 10.1109/ICRA.2017.7989046.

[39] Matthew V. Law, Nnamdi Nwagwu, Amritansh Kwatra, Seo-Young Lee, Daniel M. Diangelis, Naifang Yu, Gonzalo Gonzalez-Pumariega, Amit Rajesh, and Guy Hoffman. Affinity Diagramming with a Robot. *J. Hum.-Robot Interact.*, 13(1), March 2024. doi: 10.1145/3641514. URL https://doi.org/10.1145/3641514. Place: New York, NY, USA Publisher: Association for Computing Machinery.

[40] Jai-Yon Lee, Young Ae Song, Ji Young Jung, Hyun Jeong Kim, Bo Ram Kim, Hyun-Kyung Do, and Jae-Young Lim. Nurses' needs for care robots in integrated nursing care services. *Journal of Advanced Nursing*, 74(9):2094–2105, 2018.

[41] Wing-Yue Geoffrey Louie and Goldie Nejat. A Social Robot Learning to Facilitate an Assistive Group-Based Activity from Non-expert Caregivers. *International Journal of Social Robotics*, 12(5):1159–1176, November 2020. ISSN 1875-4805. doi: 10.1007/s12369-020-00621-4. URL https://doi.org/10.1007/s12369-020-00621-4.

[42] Michal Luria, Guy Hoffman, and Oren Zuckerman. Comparing Social Robot, Screen and Voice Interfaces for Smart-Home Control. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 580–628, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 978-1-4503-4655-9. doi: 10.1145/3025453.3025786. URL https://doi.org/10.1145/3025453.3025786. event-place: Denver, Colorado, USA.

[43] Nobuto Matsuhira, Junko Hirokawa, Hideki Ogawa, and Tatsuya Wada. Universal Design with Robots Toward the Wide Use of Robots in Daily Life Environment. In *Advances in Service Robotics*. July 2008. ISBN 978-953-7619-02-2. doi: 10.5772/5941.

[44] Meredith Mealer, Ellen L. Burnham, Colleen J. Goode, Barbara Rothbaum, and Marc Moss. The prevalence and impact of post traumatic stress disorder and burnout syndrome in nurses. *Depression and Anxiety*, 26(12):1118–1126, 2009. ISSN 1520-6394. doi: 10.1002/da.20631.

[45] Nina Moorman, Nakul Gopalan, Aman Singh, Erin Hedlund-Botti, Mariah Schrum, Chuxuan Yang, Lakshmi Seelam, and Matthew C Gombolay. Investigating the impact of experience on a user's ability to perform hierarchical abstraction. In *Robotics: Sience and Systems*, 2023.

[46] Jakob Nielsen. *Usability engineering*. Morgan Kaufmann, 1994.

[47] Jakob Nielsen. 10 usability heuristics for user interface design, 2020. URL https://www.nngroup.com/articles/ten-usability-heuristics/.

[48] Edwin Olson. Apriltag: A robust and flexible visual fiducial system. In *2011 IEEE International Conference on Robotics and Automation*, pages 3400–3407, 2011. doi: 10.1109/ICRA.2011.5979561.

[49] Aznoora Osman, Hanif Baharin, Mohammad Hafiz Ismail, and Kamaruzaman Jusoff. Paper prototyping as a rapid participatory design technique. *Computer and Information Science*, 2, 07 2009. doi: 10.5539/cis.v2n3p53.

[50] Anastasia K. Ostrowski, Cynthia Breazeal, and Hae Won Park. Long-term co-design guidelines: Empowering older adults as co-designers of social robots. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 1165–1172, 2021. doi: 10.1109/RO-MAN50785.2021.9515559.

[51] Xanthi S. Papageorgiou, Athanasios C. Dometios, and Costas S. Tzafestas. Towards a User Adaptive Assistive Robot: Learning from Demonstration Using Navigation Functions. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 965–970, 2021. doi: 10.1109/IROS51168.2021.9636200.

[52] Alexandros Paraschos, Christian Daniel, Jan R Peters, and Gerhard Neumann. Probabilistic movement primitives. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/e53a0a2978c28872a4505bdb51db06dc-Paper.pdf.

[53] Alexandros Paraschos, Christian Daniel, Jan Peters, and Gerhard Neumann. Using probabilistic movement primitives in robotics. *Autonomous Robots*, 42:529–551, 2018.

[54] Chris Paxton, Felix Jonathan, Andrew Hundt, Bilge Mutlu, and Gregory D. Hager. Evaluating Methods for End-User Creation of Robot Task Plans. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6086–6092, 2018. doi: 10.1109/IROS.2018.8594127.

[55] Maribel Pino, Mélodie Boulay, François Jouen, and Anne Rigaud. "Are we ready for robots that care for us?" Attitudes and opinions of older adults toward socially assistive robots. *Frontiers in Aging Neuroscience*, 7, 2015. ISSN 1663-4365. URL https://www.frontiersin.org/article/10.3389/fnagi.2015.00141.

[56] Natasha Randall and Selma Sabanovic. A picture might be worth a thousand words, but it's not always enough to evaluate robots. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-*

Robot Interaction, HRI '23, page 437–445, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399647. doi: 10.1145/3568162.3576970. URL https://doi.org/10.1145/3568162.3576970.

[57] Harish Ravichandar, Athanasios S. Polydoros, Sonia Chernova, and Aude Billard. Recent Advances in Robot Learning from Demonstration. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1): 297–330, May 2020. ISSN 2573-5144. doi: 10.1146/annurev-control-100819-063206. URL https://doi.org/10.1146/annurev-control-100819-063206. Publisher: Annual Reviews.

[58] Lea Rollova, Peter Hubinský, and Natália Bošková Filová. Universal design and social care: Assistive robots as other users of the built environment? *Architecture Papers of the Faculty of Architecture and Design STU*, 28:10–17, September 2023. doi: 10.2478/alfa-2023-0015.

[59] Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. No-Regret Reductions for Imitation Learning and Structured Prediction. In *14th Internation Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, FL, 2011. URL https://arxiv.org/abs/1011.0686v3.

[60] Leonel Rozo and Vedant Dave. Orientation probabilistic movement primitives on riemannian manifolds. In *5th Annual Conference on Robot Learning*, 2021. URL https://openreview.net/forum?id=csMg2h_LR37.

[61] Ljilja Ruzic, Seunghyun Tina Lee, Yilin Elaine Liu, and Jon A Sanford. Development of universal design mobile interface guidelines (udmig) for aging population. In *Universal Access in Human-Computer Interaction. Methods, Techniques, and Best Practices: 10th International Conference, UAHCI 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17-22, 2016, Proceedings, Part I 10*, pages 98–108. Springer, 2016.

[62] Joe Saunders, Dag Sverre Syrdal, Kheng Lee Koay, Nathan Burke, and Kerstin Dautenhahn. "Teach Me–Show Me"—End-User Personalization of a Smart Home and Companion Robot. *IEEE Transactions on Human-Machine Systems*, 46(1):27–40, 2016. doi: 10.1109/THMS.2015.2445105.

[63] Mariah Schrum, Muyleng Ghuy, Erin Hedlund-botti, Manisha Natarajan, Michael Johnson, and Matthew Gombolay. Concerning trends in likert scale usage in human-robot interaction: Towards improving best practices. *J. Hum.-Robot Interact.*, 12(3), apr 2023. doi: 10.1145/3572784. URL https://doi.org/10.1145/3572784.

[64] Mariah L. Schrum, Erin Hedlund-Botti, and Matthew Gombolay. Reciprocal MIND MELD: Improving Learning From Demonstration via Personalized, Reciprocal Teaching. In *6th Annual Conference on Robot Learning*, 2022.

[65] Lakshmi Seelam, Erin Hedlund-Botti, Chuxuan Yang, and Matthew Gombolay. Interface Design for Learning from Demonstration with Older Adults. In *Association for the Advancement of Artificial Intelligence Fall Symposium Series*, 2023.

[66] Soonhwa Seok, Boaventura DaCosta, and Russ Hodges. A systematic review of empirically based universal design for learning: Implementation and effectiveness of universal design in education for students with and without disabilities at the postsecondary level. *Open J. Soc. Sci.*, 06(05):171–189, 2018.

[67] Andrew Silva, Nina Moorman, William Silva, Zulfiqar Zaidi, Nakul Gopalan, and Matthew Gombolay. Lanconlearn: Learning with language to enable generalization in multi-task manipulation. *IEEE Robotics and Automation Letters*, 7(2):1635–1642, 2022. doi: 10.1109/LRA.2021.3139667.

[68] Cory-Ann Smarr, Tracy L. Mitzner, Jenay M. Beer, Akanksha Prakash, Tiffany L. Chen, Charles C. Kemp, and Wendy A. Rogers. Domestic Robots for Older Adults: Attitudes, Preferences, and Potential. *International journal of social robotics*, 6(2):229–247, April 2014. ISSN 1875-4791 1875-4805. doi: 10.1007/s12369-013-0220-0.

[69] Constantine Stephanidis, Demosthenes Akoumianakis, and Anthony Savidis. Universal design in human-computer interaction. *book: International Encyclopedia of Ergonomics and Human Factors*, 1:741–745, 2001.

[70] Simon Stepputtis, Joseph Campbell, Mariano J. Phielipp, Stefan Lee, Chitta Baral, and Heni Ben Amor. Language-conditioned imitation learning for robot manipulation tasks. *CoRR*, abs/2010.12083, 2020. URL https://arxiv.org/abs/2010.12083.

[71] Jaeyong Sung, Seok Jin, Ian Lenz, and Ashutosh Saxena. Robobarista: Learning to Manipulate Novel Objects via Deep Multimodal Embedding. January 2016.

[72] Dag Sverre Syrdal, Kerstin Dautenhahn, Kheng Koay, and Michael Walters. *The Negative Attitudes towards Robots Scale and Reactions to Robot Behaviour in a Live Human-Robot Interaction Study*. January 2009.

[73] Roger Andre Søraa, Gunhild Tøndel, Mark W. Kharas, and J Artur Serrano. What do Older Adults Want from Social Robots? A Qualitative Research Approach to Human-Robot Interaction (HRI) Studies. *International Journal of Social Robotics*, 15(3):411–424, March 2023. ISSN 1875-4805. doi: 10.1007/s12369-022-00914-w. URL https://doi.org/10.1007/s12369-022-00914-w.

[74] Adriana Tapus, Maja J. Mataric, and Brian Scassellati. Socially assistive robotics [Grand Challenges of Robotics]. *IEEE Robotics & Automation Magazine*, 14(1):35–42, March 2007. ISSN 1558-223X. doi: 10.1109/MRA.2007.339605. Conference Name: IEEE Robotics & Automation Magazine.

[75] Aleš Ude, Bojan Nemec, Tadej Petrić, and Jun Morimoto. Orientation in cartesian space dynamic movement primitives. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2997–3004, 2014. doi: 10.1109/ICRA.2014.6907291.

[76] Garrett Wilson, Christopher Pereyda, Nisha Raghunath,

Gabriel de la Cruz, Shivam Goel, Sepehr Nesaei, Bryan Minor, Maureen Schmitter-Edgecombe, Matthew E. Taylor, and Diane J. Cook. Robot-enabled support of daily activities in smart home environments. *Cognitive Systems Research*, 54:258–272, 2019. ISSN 1389-0417. doi: https://doi.org/10.1016/j.cogsys.2018.10.032. URL https://www.sciencedirect.com/science/article/pii/S1389041718302651.

[77] Selma Šabanović, Wan-Ling Chang, Casey C. Bennett, Jennifer A. Piatt, and David Hakken. A Robot of My Own: Participatory Design of Socially Assistive Robots for Independently Living Older Adults Diagnosed with Depression. In Jia Zhou and Gavriel Salvendy, editors, *Human Aspects of IT for the Aged Population. Design for Aging*, pages 104–114, Cham, 2015. Springer International Publishing. ISBN 978-3-319-20892-3.