Design and Evaluation of AR-Based Real-Time Feedback System for Kinesthetic Robot Teaching

Muhammad Bilal
School of Computing and Information Systems, The University of Melbourne
Melbourne, Australia
m.bilal@unimelb.edu.au

Tharaka Sachintha Ratnayake
School of Computing and Information Systems, The University of Melbourne
Melbourne, Australia
ratnayakemt@unimelb.edu.au

D. Antony Chacon
School of Computing and Information Systems, The University of Melbourne
Melbourne, Australia
antony.chacon@unimelb.edu.au

Nir Lipovetzky
School of Computing and Information Systems, The University of Melbourne
Melbourne, Australia
nir.lipovetzky@unimelb.edu.au

Denny Oetomo
Department of Mechanical Engineering, The University of Melbourne
Melbourne, Australia
doetomo@unimelb.edu.au

Wafa Johal
School of Computing and Information Systems, The University of Melbourne
Melbourne, Australia
wafa.johal@unimelb.edu.au

Abstract

Learning from Demonstration (LfD) allows novice users to teach robots through demonstrations without coding; however, such demonstrations are often suboptimal and can limit robot performance. To better support novices, we investigate the design of a feedback system that enables effective human-robot communication during demonstrations. We first conducted a focus group study ($N = 9$) to identify effective ways of visualizing key robot information, including joint limits, self-collisions, and manipulability. Guided by these insights, we designed an AR-based real-time feedback system and evaluated it in a between-subjects user study ($N = 36$) on a 7-DoF collaborative robot. Participants performed two tasks—insertion and pouring—with the second task enabling assessment of participants' learning across tasks. Results show that real-time feedback reduced demonstration time, increased task completion rate, lowered perceived mental workload, and improved adherence to robot kinematic constraints. These findings demonstrate the effectiveness of the real-time feedback system for intuitive and effective robot teaching.

CCS Concepts

• **Human-centered computing** → **Mixed / augmented reality**; **User studies**; • **Computing methodologies** → **Learning from demonstrations**.

Keywords

Human-Robot Interaction, Learning from Demonstration, Kinesthetic Teaching, Augmented Reality

ACM Reference Format:
Muhammad Bilal, Tharaka Sachintha Ratnayake, D. Antony Chacon, Nir Lipovetzky, Denny Oetomo, and Wafa Johal. 2026. Design and Evaluation of AR-Based Real-Time Feedback System for Kinesthetic Robot Teaching. In *Designing Interactive Systems Conference (DIS '26), June 13–17, 2026, Singapore, Singapore.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3800645.3812951

DIS '26, Singapore, Singapore

ACM ISBN 979-8-4007-2563-0/2026/06
https://doi.org/10.1145/3800645.3812951

1 Introduction

Collaborative robots (cobots) are becoming increasingly capable, with advances in sensing, control, and computation enabling them to perform a growing range of tasks while simultaneously decreasing in size and cost and improving in robustness [32, 72]. Despite these advances, the adoption of cobots in many workplaces, such as SMEs, remains limited [38]. In addition, we observe persistent concerns that automation may displace human workers [1, 27, 31] and negatively impact the quality of work [38, 59]. Such concerns are often exacerbated by the limited accessibility of robotic systems for novice users, as most robots still require specialized technical expertise to be programmed and operated [5, 7, 38]. From a worker-centered design perspective, improving the democratization of robotic technologies is therefore not only a technical challenge but also a means of empowering workers to appropriate collaborative robots as flexible tools that support their own ways of working, rather than having their practices constrained by the technology [30]. In response, prior work has explored interaction paradigms that enable non-expert users to engage with robotic systems in more natural and intuitive ways [46, 80]. One of the most widely adopted non-technical approaches is Learning from Demonstration (LfD) [15, 62]. In kinesthetic teaching, a common modality within LfD, users teach a robotic system by physically guiding its motion—moving its joints to show how a task should be performed. Rather than writing complex code, users convey their intent through action and embodied interaction. By reducing reliance on traditional programming, LfD has been shown to lower barriers to entry and broaden participation in the adaptation and deployment of robotic systems [18, 62, 78].

Teaching a collaborative robot is often more difficult than simply showing it what to do. In practice, novice users frequently provide demonstrations that are incomplete, inconsistent, or poorly suited to the robot's learning process [6, 23]. Traditional approaches tend

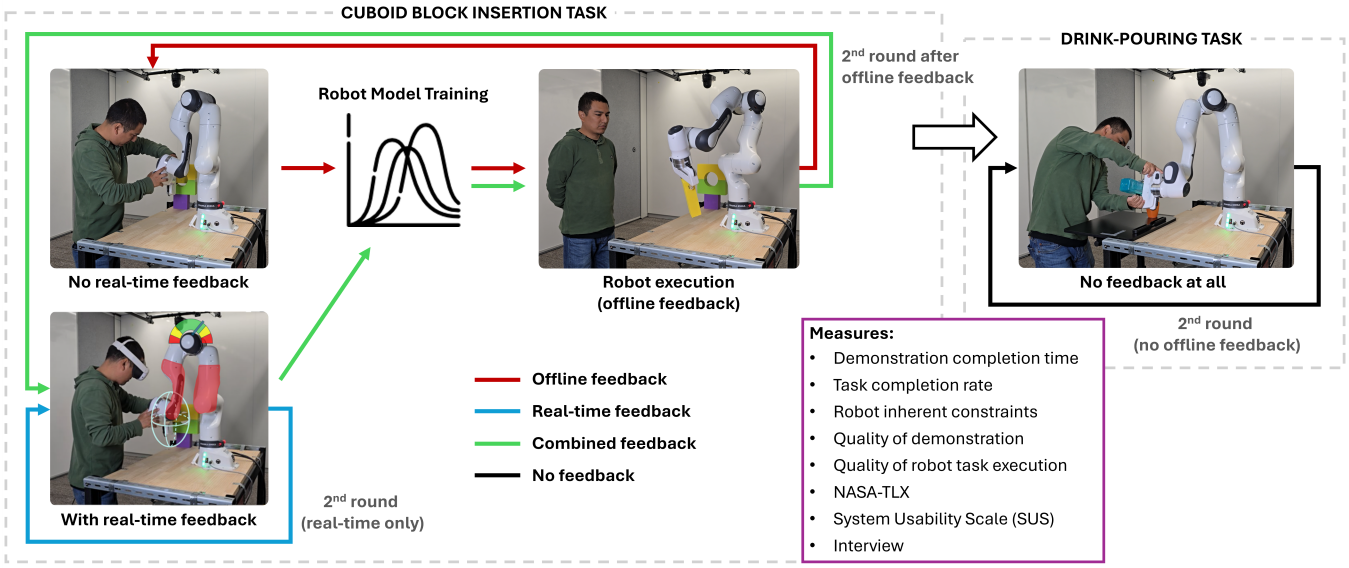


Figure 1: Overview: The study compares three feedback conditions during kinesthetic robot teaching: offline feedback (red), real-time feedback (blue), and combined feedback (green). Participants first performed a cuboid block insertion task by providing three kinesthetic demonstrations to a 7-DoF collaborative robot, followed by a second demonstration round depending on the feedback condition. A second drink-pouring task was then performed with no feedback for any group, enabling assessment of participants' learning across tasks. AR visualizations are illustrative only.

to mitigate this issue by ignoring or filtering out "bad" demonstrations, without addressing why these breakdowns occur in the first place [33, 36]. A key reason is that human teachers often struggle to understand what the robot has actually learned at any given moment. Unlike human students, robots do not naturally signal confusion, uncertainty, or partial understanding [67]. As a result, people rely on assumptions about the robot's thinking that are often inaccurate, leading to frustration and suboptimal teaching.

This challenge is rooted in the opaque nature of the robot's learning process. When teaching other humans, people adjust their instruction based on visible cues such as questions, hesitation, or mistakes. When teaching robots, these cues are largely absent, turning teaching into a one-sided activity rather than a mutual exchange. To address this gap, robot learning must be treated as a dialogue in which the robot actively communicates its current state to the teacher through clear and accessible feedback, such as visual, haptic, or auditory signals [40, 68, 73]. Prior work suggests that when learning and communication are tightly connected in this way, teaching becomes more effective, trust is strengthened, and both humans and robots can better adapt to one another [34, 45, 64, 76]. From a design perspective, this highlights the need for new communication interfaces that help robot "students" express what they understand, what they do not understand, and when they need help, making teaching a shared and responsive process rather than a guessing game [3, 4, 24, 46].

Augmented reality (AR) technology has emerged as a significant advancement in human-robot interaction (HRI), as it enables the integration of three-dimensional virtual graphics directly into the user's real-world field of view [39, 84, 89]. By visually enriching the physical environment, AR not only improves the user's perceptual awareness of robotic systems but also provides intuitive spatial feedback during interaction. Building on these capabilities, AR-based tools have been successfully deployed in industrial environments, where they help reduce cognitive load, facilitate robot training, and support calibration and inspection processes for complex assembly tasks [50, 52, 53].

In this work, we address the overarching question: *How can robots communicate internal constraints to users in human-robot Learning from Demonstration interactions?* To explore this question, we conducted a focus group ($N = 9$) with extended reality (XR) developers and roboticists as design stakeholders to generate interaction concepts for communicating a robot's internal constraints (e.g., joint limits, potential self-collisions, and singular configurations). We derive design findings that surface opportunities and tensions around safety, alerting mechanisms, cognitive load, and opportunities for customization.

Building on these insights, we developed an AR-based feedback system to make robot constraints perceptible in real time. To evaluate its effectiveness, we conducted a two-task between-subjects user study ($N = 36$). Our findings show that novice users who received real-time AR feedback not only performed better on the initial task but also retained these benefits when the feedback was removed in the second task, demonstrating measurable learning. Moreover, the feedback supported users in developing a better understanding of the robot's motion constraints, leading to improved robot task performance even without visual assistance. Figure 1 illustrates an overview of the experimental study.

2 Related Work

One of the key goals of LfD is to enable novice users to efficiently and effectively demonstrate tasks to robotic systems [8, 62]. A growing body of research has focused on understanding and supporting the human teaching process, including ways to measure, evaluate, and improve user demonstrations [13, 61, 66, 69]. However, most existing approaches rely on *offline* feedback, where users observe the robot's reproduced motion after completing demonstrations in order to refine or correct their guidance [2, 64, 76]. While such feedback can improve robot learning outcomes, it does not provide guidance during the demonstration itself, highlighting an opportunity to explore *real-time feedback mechanisms*, such as those leveraging AR technology, to support users during task demonstration.

2.1 Robot Teaching Interfaces

Kinesthetic teaching allows users to guide a robot through physical interaction, offering an intuitive starting point for demonstration-based learning [14, 62]. However, this interaction becomes challenging when robots have many joints or operate under physical constraints [37]. During teaching, users must implicitly account for issues such as limited joint ranges, complex postures, or impending collisions, even though these factors are often invisible or difficult to perceive through touch alone [12]. As a result, demonstrators may unintentionally guide the robot into problematic configurations, producing demonstrations that are difficult for the robot to interpret or reproduce [12, 37, 79].

Augmented reality offers an opportunity to address these challenges by making aspects of the robot's internal state visible during teaching [22]. When AR cues are anchored to the robot, they can communicate information such as approaching limits, constrained motion, or regions of reduced flexibility directly within the user's field of view [84]. This added transparency can help users better understand how the robot experiences the task, enabling more informed and effective demonstrations. At the same time, careful design is required to avoid overwhelming the demonstrator, as visualizing too many signals at once may increase cognitive load, thereby distracting users from the physical task and reducing teaching quality [22, 25]. From a human-centered perspective, it is therefore important not only to support successful task demonstrations but also to consider user-focused measures such as perceived mental workload and the ease of providing demonstrations without violating the robot's physical constraints.

2.2 AR for Teaching Robots

Prior work has explored AR systems in human-robot interaction, with most applications focusing on robot teleoperation or programming interfaces rather than supporting users during kinesthetic demonstrations [20, 85, 91, 92]. For instance, AR has been used to specify goal poses, preview robot motion, and define task trajectories through virtual overlays or haptic input devices [55, 63]. Other studies have investigated tablet- or headset-based AR interfaces for commanding or programming robots, reporting effects on user workload and task efficiency depending on the task and platform [20, 21, 56]. By providing spatially aligned visual feedback within the user's field of view, AR allows users to perceive robot motion, limitations, and task-relevant information without interrupting the flow of the demonstrations [39, 84]. Recently, researchers have investigated how additional sensory cues—including haptic feedback and AR visualizations—can improve user awareness, guidance, and performance during human-robot interaction [34, 75]. However, these AR-based systems have focused on remote interaction with robots, with comparatively less attention given to examining how these design principles translate to hands-on kinesthetic teaching with physical robots [85, 89], which enables robots to learn more effectively and execute tasks more successfully [13, 65].

Despite the promise of augmented reality for making robot status visible during interaction, there is still limited understanding of how AR-based feedback can be integrated into kinesthetic teaching in a way that meaningfully supports users without overloading them [20, 22, 85]. In particular, little is known about how such feedback can help users navigate constrained teaching situations—such as avoiding collisions, respecting motion limits, or steering clear of problematic postures—while maintaining a smooth and efficient teaching experience. This study builds on prior work by designing and evaluating an AR-based feedback system that communicates robot status in real time during kinesthetic demonstrations. Rather than treating AR merely as a programming aid, we examine how in-situ visual feedback supports users during teaching, how it shapes demonstration quality and efficiency, and how it compares with more conventional forms of offline or post-hoc feedback. To this end, we pose the following research questions:

RQ1: What design considerations and trade-offs emerge when envisioning the real-time communication of a robot's internal states during LfD?

RQ2: How does the presence of AR-based constraint feedback influence how users demonstrate tasks to a robot, including their effort, strategies, and accommodation of robot constraints?

RQ3: How does AR-based constraint feedback influence a robot's ability to learn from demonstrations and successfully execute the demonstrated tasks?

3 Study 1: Focus Group

To explore the design space for communication during kinesthetic demonstrations, we conducted an in-person focus group with roboticists and AR/VR developers. This mix enabled insights into both robot capabilities and AR/VR interaction possibilities. Participants were paired to ensure that each group combined both areas of expertise, thereby supporting cross-domain discussions [11]. Through this focus group, we address **RQ1**.

3.1 Participants Recruitment

Nine participants (1 woman, 8 men; M_{age} = 28 years, SD_{age} = 2.82) were recruited for the study (see Table 1). Their expertise spanned AR/VR development, robot teleoperation, manipulation, exoskeleton control, XR research, physics-based simulation, and probabilistic robotics, with 1 to 8 years of experience in AR/VR, robotics, or both.

Table 1: Participant Demographic Information

ID	Gender	Age	Experience Type	Yrs. of Exp.	Expertise Description
P1G1	Male	29	AR/VR and Robotics	8	Developed AR applications; XR/AR PhD; robotics project experience since undergrad.
P2G1	Female	28	AR/VR and Robotics	7	PhD in Mechanical Engineering and HRI; experience in VR data processing, robot building, probabilistic robotics, and HCI.
P3G1	Male	26	Robotics	3	Worked on bipedal robots, exoskeletons, and manipulators for HRI.
P4G2	Male	33	AR/VR and Robotics	2	Researched robot teleoperation using AR glasses.
P5G2	Male	30	AR/VR	2	Developed AR/VR applications for architectural heritage reconstruction.
P6G2	Male	28	AR/VR	1	Worked on physics simulation visualizations in VR.
P7G3	Male	23	AR/VR	1	Conducted XR research.
P8G3	Male	29	AR/VR	3	Experience in motion synthesis and soft-rigid body simulation.
P9G3	Male	26	AR/VR and Robotics	6	VR scene design; 4+ years working with NAO and TIAGo robots.

3.2 Procedure

The study was a structured, multi-stage session conducted in our School's User Experience Lab. Participants provided consent, reviewed a Plain Language Statement, completed a demographic questionnaire, and then took part in a brief icebreaker. Afterward, they were assigned to groups of three and introduced to LfD. The facilitator outlined key concepts, demonstration modalities, and common challenges to establish a shared understanding before ideation. Participants then engaged in a hands-on session with a Franka Emika 7-DoF robot[1], exploring predefined interactions to understand its capabilities and limitations.

After the exploration phase, participants completed a structured ideation activity based on rapid brainstorming (Crazy 4s) [26]. Focusing on three robot constraints—joint limits, self-collision, and manipulability—each participant generated four AR visualization and interaction ideas per round. They then regrouped to present, discuss, and collaboratively select two ideas, combining elements where relevant. This iterative divergence-convergence process supported both broad exploration and refinement. After three rounds, the groups pitched their top ideas, followed by a cross-group discussion. Participants could then refine their concepts before voting on the strongest idea for each feature based on clarity, usefulness, and support for AR-mediated robot learning.

3.3 Analysis

Audio recordings were transcribed using Otter.ai[2], then reviewed and corrected by the first author. Identifiable information was removed and replaced with pseudonyms (e.g., P1G1). Transcripts and drawings were imported into a shared Miro[3] board for analysis. Two coders independently familiarized themselves with the data and generated inductive codes [77]. They first aligned codes through a line-by-line review of Group 1, then jointly coded Group 2, refining codes and emerging themes. All code iterations were systematically documented in Miro for version control. Finally, coders developed and refined themes through discussion. The second author verified all reported quotes against the original audio before the recordings were deleted.

[1]https://franka.de/franka-research-3-arm
[2]https://otter.ai
[3]https://miro.com

3.4 Findings

In this section, we present the findings from the focus group, organized into three holistic categories based on joint limits, self-collision and manipulability. Figure 2 presents the final sketches that informed the features implemented in the AR-based feedback system used in the subsequent user study. Additional relevant drawings produced by participants are provided in Appendix A.4.

3.4.1 Joint Limits. Our focus group findings show that novices struggle with joint limits, as these constraints are rarely made explicit in LfD. Users must often infer these limits through failures (e.g., the robot stopping) while focusing on task goals rather than joint angles, making it difficult to identify the cause. This mismatch between users' mental models and hidden constraints highlights the need for clearer communication of joint-specific limits. We next present the focus group insights, organized in a bottom-up hierarchy (see Figure A1).

Color-Based Awareness of Joint Limits. Participants emphasized immediate, intuitive visual indicators for joint limits, with color as the preferred modality. Many favored a familiar stoplight scheme—green for safe and red for unsafe—due to its low cognitive demand (e.g., "*If it is okay they're green, if they are not okay they become red*" (P9G3)). They also highlighted the importance of progressive feedback as limits are approached. Suggestions included color transitions or expanding cues to convey increasing urgency, such as a widening, blinking red indicator: "*when I turn further, that red color strip will broaden up further and then like it will start blinking*

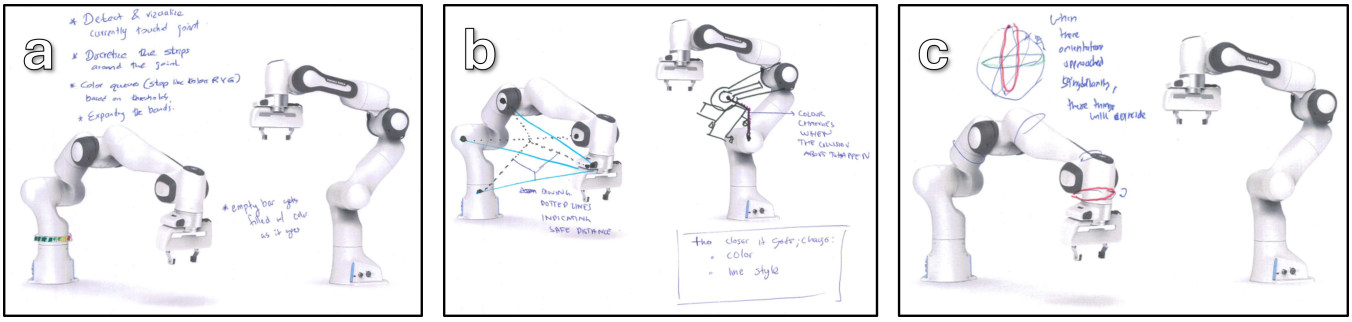


Figure 2: Final visualization designs of each feature from the focus group study: (a) joint limits, (b) self-collision, and (c) manipulability.

[showing] you have reached the limit" (P1G1), or smooth gradients to provide finer-grained feedback for expert users: "*green means that it's good and then it might change to like a yellow and then an orange and then a red and then it's like a smoother transition*" (P2G1).

Quantifying Proximity to Joint Limits. Beyond basic color cues, participants emphasized the need for explicit quantification of joint-limit proximity and direction, especially for precise or incremental tasks. Discretized visuals were favored because they provide an immediate sense of remaining range without mental calculation: "*Discretized version would be better because it would give a sense of ... how much you've got to go*" (P3G1). Movement-aligned cues, such as arrows that progressively change color, were seen as intuitive for mapping proximity: *once you get closer to a particular limit in one direction, you can have the parts of the arrows ... more red ... and if it's far from the joint limits ... more green*" (P6G2). Participants also highlighted numerical information for expert users who require high precision, while noting that it could overwhelm novices: "*present the actual information to the user so like the numerical rotation degrees ... they can actually see if it's close to it or not*" (P6G2).

Integrated and Dynamic Feedback. Participants emphasized that joint-limit feedback should remain spatially grounded in the XR scene to minimize cognitive load. They favored visual overlays co-located with the robot—such as rings, bars, or annotations on or near the joints—for intuitive, non-disruptive interaction: "*Since the joints are circular, I was thinking of putting rings, and whenever they get close to the limit, they become a different color and the bars move*" (P4G2). To reduce clutter on robots with many degrees of freedom, participants recommended context-sensitive displays showing only the actively manipulated joint. As one participant suggested, *How about like AR automatically detects the joint that we are rotating*" (P1G1). Selective disclosure simplifies the interface while still providing critical feedback: *So the people can see this ... this is if you want to hide all of the complexities ... from the users*" (P6G2). While localized feedback was preferred during manipulation, participants acknowledged the value of optional global summaries—such as side panels or dashboards—to give an overview without competing for attention: "*We can actually have a list of ... ideas next to it in summary ... even if we don't draw them exactly how we want*" (P2G1).

Escalating Alerts Near Critical Limits. Participants emphasized feedback escalation that is clear yet non-intrusive, as users must react quickly near joint limits. Early warnings should be subtle, using gradual visual changes to signal risk without disrupting the task: "*as you reach like a problem state, the band becomes more prominent ... you are in the red zone*" (P2G1). This approach allows users to adjust their motion proactively while maintaining focus on the manipulation. More disruptive cues, like blinking or rapid animation, were reserved for critical thresholds when a hard limit is imminent: "*I do like blinking in the sense that ... it causes more awareness ... probably something wrong*" (P3G1); "*when I reach the limit it starts blinking rapidly as well*" (P1G1). Some participants suggested optional auditory cues as a complementary channel for visually dense tasks, reinforcing visual feedback without replacing it: "*if it is going to the end it goes to red if it is okay it's green ... can also make like an audio that it is going to be limited*" (P9G3).

Customization and User-Dependent Interactions. Participants highlighted the need for customizable joint-limit visualizations to accommodate diverse expertise and task goals. Novices preferred simple, high-level cues, while experts benefited from granular, on-demand information. Interaction techniques allowing users to request detail were favored to avoid clutter: "*click on a joint and a stoplight for that joint specifically appears ... very binary*" (P2G1); "*If we could make it automatic ... pops up ... first of all it's not annoying*" (P3G1). Advanced information was recommended to remain hidden by default, surfaced only through deliberate actions like tapping or pinching a joint. This selective disclosure helps users control content and complexity: "*... tap/pinch joint ... info is not there unless you request it ... pick three and almost like-the info is not there unless you request it*" (P1G1).

3.4.2 Self-Collision. Participants reported that self-collision poses a particular challenge for novices because it lies outside their focus on the end-effector and is only revealed through execution failures. Novices often lack awareness of the robot's full-body geometry (see Figure A2).

Collision Proximity Awareness. Effective self-collision support requires immediate perception of proximity between robot components. This baseline awareness was described as essential for both novice and expert users, particularly because self-collisions can emerge gradually and unintentionally during complex demonstrations. Participants favored cues that visualize distance and urgency, such as an expanding marker: "*... it starts with a small cross, and the*

cross starts growing bigger as you move towards it ... green, and then you change the color to orange, and then to red ..." (P1G1).

Color-coded escalation was recommended, with transitions ranging from green (safe) to red (contact imminent), visually anchored on the at-risk components: "*if close to self-collision, the two limbs that are anticipated to collide will both become highlighted, and maybe with exclamation points or something as a warning*" (P2G1). Additional dynamic cues, like arrows or translucent meshes, were proposed to reinforce spatial understanding: "*... when two links are getting closer to each other, an arrow pops up ... yellow when still in transition ... red when actually collide*" (P3G1).

Directional Guidance for Recovery After Collision Avoidance. Once collision risk is detected, participants emphasized that feedback should provide actionable guidance, with directional cues complementing proximity and urgency indicators: "*... at the point that I am touching, giving me the cues ... you can move to the other side [cue]*" (P1G1). Motion or animation was suggested to enhance interpretability, with arrows that "push" in a corrective direction seen as particularly effective: "*... the arrows animated to move ... showing like, go that way, go that way ... makes it clearer*" (P2G1). Directional guidance should be integrated with color-coded urgency cues to convey both risk and corrective options: "*... point the user to which way you can correct this collision ... shows how to resolve the collision*" (P9G3).

3D Volume and Mesh Representations. Participants proposed volumetric feedback to represent collision boundaries in 3D space, such as mesh shells or translucent casings around joints, moving beyond planar cues: "*... instead of just using color, actually having almost a casing compress ... kind of like those things with the spikes ... you almost can see an extended version ...*" (P1G1). These elements were expected to behave dynamically, appearing or deforming as components approach collision. Progressive mesh visualization was seen as an intuitive way to convey risk: "*... mesh that shows around there ... if it's coming closer, the mesh starts to appear*" (P5G2); "*... basically have meshes for each joint ... whenever they get closer, they start to appear*" (P4G2).

Distinguishing Normal and Danger States. Participants emphasized distinguishing low-risk proximity from dangerous states to reduce cognitive load. During safe operation, feedback should remain subtle and unobtrusive: "*... normal situation ... nothing appears ... as you move, you're starting to see something happening ...*" (P3G1). More intrusive cues—such as strong colors, arrows, or alerts—should appear only after a risk threshold is crossed, following a clear hierarchy: "*... the dots are not meant to be a trigger ... arrows are meant to be like, hey, you need ... color scale giving inkling you might start having a problem ... here's the thing that's red*" (P2G1).

3.4.3 Manipulability. Participants found manipulability to be one of the least intuitive constraints, as it depends on the full joint configuration rather than visible factors. Novices often struggled to understand why seemingly correct poses limited motion (see Figure A3).

Global vs. Local Visualization of Singularity. Manipulability was described as a multi-scale problem requiring both global and local awareness. Singularities were framed as whole-body phenomena, motivating global feedback such as clouds or envelopes around the robot: "*... singularity is less a local joint thing, but more like a global whole robot thing. So, like a cloud around the robot or something that turns the color of the robot to red, yellow ...*" (P3G1). However, participants emphasized that global cues should not obscure the source of risk. When specific joints drive degradation, localized highlighting enables targeted correction. As one noted, "*... those actual limbs or joints ... get color highlighted ... if you're getting close to a singularity, red or yellow*" (P2G1), helping users identify and address instability during demonstration.

Manipulability Awareness Alerts. Participants emphasized making approaching manipulability limits perceptible through intuitive alerts. Color coding was the most common method, with transitions to yellow or red signaling degrading stability or risk of singularity: "*... joint gets technically disabled ... red color strip is shown, like you cannot move it further*" (P1G1). Beyond color, participants suggested augmenting feedback with subtle light effects or optional auditory cues to reinforce diminishing manipulability without overwhelming the user: "*... lights can get to the red ... also can make a sound ... slow sound, then faster ... like a car when you want to park it*" (P9G3).

Directional Guidance and Preventive Feedback. Participants noted that awareness alone is insufficient for manipulability issues; users also need guidance to avoid singularities. Prescriptive feedback was recommended: "*... if you were close to a singularity ... suggesting what to do to avoid it ... arrows could be a cool idea*" (P2G1). Directional cues, such as arrows or animations, should indicate safer movement in real time and adapt dynamically: "*... this tells you what to do to avoid [singularity] ... or prevents you from letting it happen*" (P6G2). Some suggested predictive feedback via simulation to anticipate singularities and propose alternative joint usage: "*... calculate ... simulation ... check the singularity ... if singularity occurs, notify alert ... use other joints and avoid until these joints avoid the singularity*" (P8G3).

3.5 Discussion

Based on our findings, we derive the following implications for the design of XR interfaces supporting robot learning from demonstration under kinematic constraints.

***Implication 1**: Risk and Constraints Should Be Color-Coded.* Joint limits, self-collision risk, and manipulability should be communicated using color coding. Stoplight-style mappings (green-yellow-red) should be used to indicate urgency with minimal cognitive effort. Continuous color gradients should convey proximity to constraints and support the anticipation of risk.

***Implication 2**: Awareness Cues Should Be Layered with Actionable Guidance.* Constraint feedback should follow a layered design. Color changes and visual highlights should signal the presence of risk. When thresholds are exceeded, directional and animated cues should provide explicit guidance for corrective action, such as indicating safe movement directions.

***Implication 3**: Constraint Feedback Should Escalate to Avoid Visual Overload.* Constraint feedback should remain visually minimal during safe operation and escalate only when constraints become

relevant or dangerous. Feedback should be spatially anchored to affected robot components, limited to active joints, and clearly distinguish normal from dangerous states to reduce clutter in XR environments.

***Implication 4**: Feedback Granularity Should Adapt to User Expertise.* The level of detail in constraint feedback should adapt to users' expertise and task demands. Novices should be supported with simple, binary indicators, while experts should have access to continuous gradients and numerical values. Detailed information should be revealed on demand to maintain simplicity during routine interaction.

4 Study 2: Lab Study Evaluation

Based on the design implications obtained in the focus group. we implemented a set of visualizations to assess their effect on human-robot interaction during two LfD tasks, addressing research questions **RQ2** and **RQ3**. For that, we conducted a lab study with novice demonstrators. The study was approved by our university ethics committee (Reference No. 2025-24636-71568-5).

4.1 Robot Key Features Visualization

We developed an XR application to visualize the three key components for LfD with robotic manipulators: (1) joint limits, (2) self-collision, and (3) manipulability. We used the latest Unity engine (Unity 6) and the Meta Quest 3 headset. We selected this headset for its passthrough capabilities and broad platform support, which help lower barriers to adoption and improve replicability.

The visualization design followed the implications identified in the focus group study. First, we implemented **recognizable color-coded cues**, using green–yellow–red for joint limits and self-collision, and blue–purple–red for manipulability. Second, we provided **actionable guidance**, particularly for joint limits, to help users respond appropriately. Third, to **reduce information overload** [74], visualizations are not continuously displayed; instead, each component appears only when its magnitude approaches the respective limit, with color transparency (α) increasing gradually as the limit is approached. Finally, in line with the implication that **feedback granularity should adapt to user expertise**, we designed the visualizations for simplicity, providing only the basic indicators suitable for novice users. Detailed or numerical feedback was not implemented, as our lab study focused exclusively on novice participants. Further details are provided below for each visualization component.

4.1.1 Joint Limits. We chose an arc overlay design that displays the actual angular range of motion for each joint, as shown in Figure 3. These arcs have a fixed thickness of 5 cm and use three colors to indicate proximity to the joint limits at either extreme. Angles within 10 degrees of a limit are shown in red, angles between 10 and 20 degrees are shown in orange, and angles beyond 20 degrees are shown in green, with smooth gradients between colors. Each arc includes a pointer, similar to clock hands, to indicate the current joint angle. In addition, we employed dynamic transparency: the arc is completely invisible when the joint is within a safe range (angular distance from the limit above 30 degrees) and becomes progressively more opaque as the joint approaches either limit (see Figure 3(a)–(d)).

4.1.2 Self-Collision. For self-collision visualization, we compute in real time the distances between all possible pairs of robot links and identify the pair with the smallest distance. Only this closest link pair is visualized at any given time. The two links are highlighted and connected by a dashed line to indicate their proximity. To facilitate quick identification, we display a label for each highlighted component (a number for links and RF/LF for the right and left fingers of the end effector, respectively). Similar to the joint limits visualization, we employ dynamic transparency: no visualization is shown when the links are at a safe distance from each other (greater than 20 cm). As the distance decreases, the visualization progressively becomes more opaque, transitions from green to red, and the connecting line changes from dashed to solid, as shown in Figure 4.

4.1.3 Manipulability. We visualize robot manipulability using a manipulability ellipsoid, which represents the set of achievable end-effector velocities for unit-norm joint velocities at a given configuration [90]. The ellipsoid's principal axes correspond to the singular vectors of the Jacobian, with axis lengths proportional to the associated singular values [28]; larger volumes indicate higher manipulability, while smaller, flattened ellipsoids indicate reduced manipulability. The ellipsoid is rendered at the end effector to provide immediate feedback during execution. Its color transitions from blue to red as manipulability decreases, with transparency varying between 0 and 0.5 (see Figure 5). Ellipsoids exceeding a volume threshold of 0.2 cm^3 are hidden to reduce clutter. Figure 6 shows all visualization components simultaneously, including joint-limit and self-collision indicators, along with a QR-based anchor at the robot base to ensure accurate AR alignment.

4.2 Experimental Setup

The study took place in a lab environment equipped with the Franka Emika 7-DoF Panda robot, a collaborative robotic manipulator. This robot arm is equipped with torque sensors in all 7 joints, making it safe and suitable for human-robot interaction applications. We used the default gravity compensation controller to counteract the robot's inherent weight, allowing participants to kinesthetically manipulate the robot's joints without additional force.

For the XR rendering, we employed the Meta Quest 3 as the head-mounted display (HMD) to visualize key robot features using its passthrough capabilities.

4.3 Experimental Conditions

This study investigated whether augmenting demonstrations with AR-based feedback improves user performance compared to traditional offline feedback, whether real-time AR feedback alone provides measurable benefit, and whether combining real-time and offline feedback further enhances interaction. We designed three conditions:

Offline Feedback — Participants demonstrated tasks without real-time guidance. After the first three demonstrations, they observed the robot's reproduced execution and used this offline feedback to adjust subsequent demonstrations.

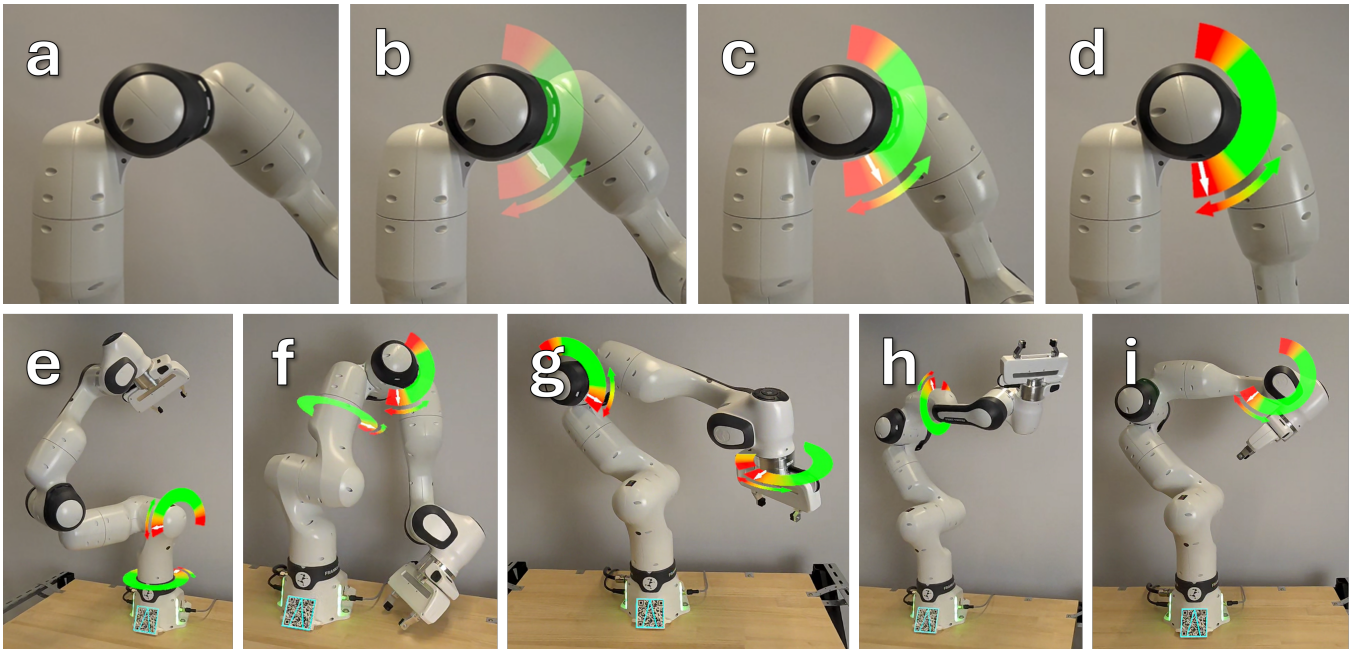


Figure 3: First row: joint-limit visualization on joint 4 based on the current absolute angular distance to the joint limit (δ): (a) no visualization, $\delta > 30°$, $\alpha = 0$; (b) $20 < \delta < 30°$, $0 < \alpha < 0.5$; (c) $10 < \delta < 20°$, $0.5 < \alpha < 1$; and (d) $\delta < 10°$, $\alpha = 1$. Second row: joint-limit visualizations on the seven joints (q_1 to q_7): (e) q_1 and q_2; (f) q_3 and q_4; (g) q_4 and q_7; (h) q_5; and (i) q_6.

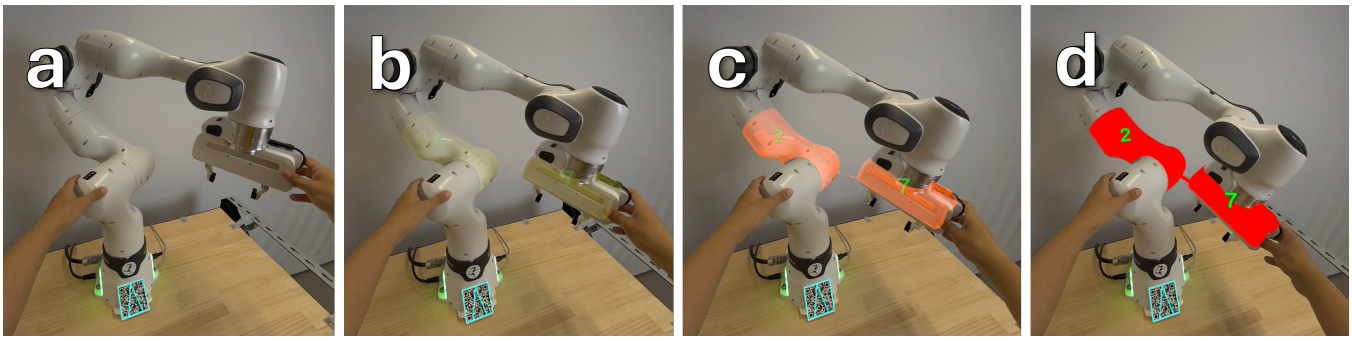


Figure 4: Self-collision visualizations between links 2 and 7 based on the surface-to-surface distance between them: (a) very safe distance above 10 cm (no visualization), (b) safe distance below 10 cm with links highlighted in green with high transparency, (c) warning distance with links in orange, and (d) danger distance with links in solid red.

Real-Time Feedback — Participants received continuous AR feedback during demonstrations, showing key task-relevant states to guide movements in real time. No execution playback was provided between rounds.

Combined Feedback — Participants received both real-time AR cues and offline execution playback, enabling immediate guidance during demonstrations and reflection on the robot's reproduced motion afterward.

4.4 Study Goals

Study 2 builds on the design insights identified in Study 1. Specifically, the study investigates how real-time, AR-based constraint feedback derived from focus group findings on awareness, escalation, and actionable guidance influences key aspects of interaction during LfD. We examine how different feedback modalities shape (1) the quality and efficiency of human demonstrations, (2) how users accommodate inherent robot constraints such as joint limits, self-collisions, and reduced manipulability, and (3) the resulting quality and reliability of robot task execution. In addition, we assess how these feedback designs affect users' subjective experience, including perceived workload and system usability.

By comparing offline, real-time, and combined feedback conditions across tasks of varying complexity, the study aims to understand not only whether real-time feedback is beneficial, but how and under what interactional circumstances it supports learning,

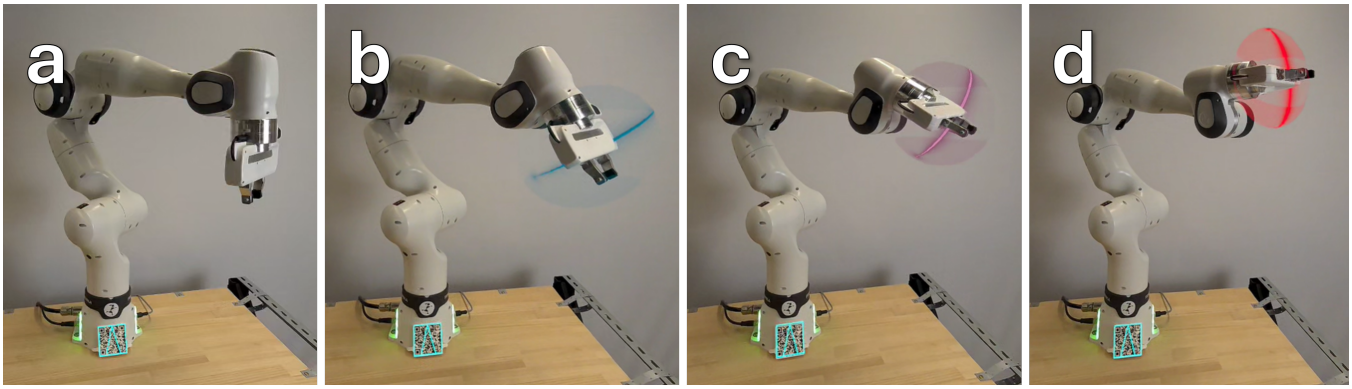


Figure 5: Manipulability ellipsoid anchored to the robot's end effector: (a) high manipulability, represented by a large ellipsoid (not visualized); (b) medium manipulability, with reduced volume shown in blue; (c) low manipulability, with further volume reduction indicated in purple (blue–red transition); and (d) very low manipulability, represented by a small ellipsoid shown in red.

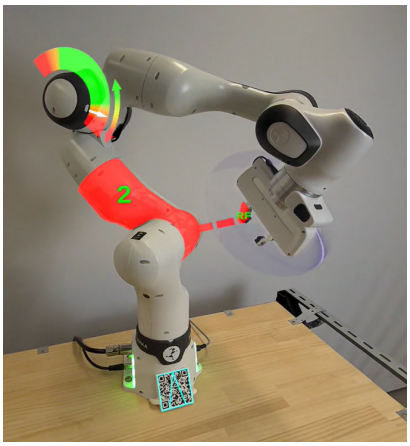


Figure 6: All three visualization features presented simultaneously on the robot: joint limit on joint 4, self-collision between link 2 and the right finger (RF), and a manipulability spheroid on the end effector. QR-code-based spatial anchor placed on link 0.

confidence, and effective teaching. This exploratory framing allows us to surface patterns and design implications that inform the role of feedback timing and modality as interaction design variables in embodied robot teaching.

4.5 Experimental Tasks

We evaluated two manipulation tasks—cuboid block insertion and drink pouring—commonly used in prior work [22, 82] for their generic nature and lack of required domain-specific expertise.

4.5.1 Insertion Task. Participants taught the robot to pick up a cuboid block ($5 \times 5 \times 20$ cm) and partially insert it into a 10 cm-diameter hole, as shown in Figure 7(a). The task required avoiding collisions with the hole edges and the robot itself, while aligning the block perpendicularly to the hole's surface.

4.5.2 Pouring Task. Participants taught the robot to grasp a bottle and move it toward a cup to simulate pouring, as seen in Figure 7(b). Success required careful joint coordination to maintain bottle orientation, avoid collisions, and ensure a natural pouring trajectory. Real liquid was not used for safety, given the novice participants and sensitive equipment.

The study used a between-subjects design in which participants were assigned to one feedback condition. Each participant provided multiple demonstrations over time, allowing us to account for changes across demonstrations while controlling for individual differences between participants.

4.6 Participants Recruitment

We conducted a priori power analysis using G*Power [29] to determine the required sample size, after which we recruited 36 participants (15 women, 21 men; M_{age} = 26.34 years, SD_{age} = 4.02) through social media advertisements and campus notice boards. To maintain consistency across experimental conditions, participants were evenly distributed such that each group comprised 7 males and 5 females (i.e., 12 participants per group). In addition to demographic questions, we used the questionnaire from [51] to ensure that all participants were novice users. Following guidance from our university ethics committee and to avoid any risk to participants, we also included two screening questions: *Do you feel uncomfortable or scared when you are around robots?* and *Do you experience motion sickness when using Augmented Reality (AR) devices?* A positive answer to either question would make the person ineligible to participate. Moreover, none of the participants who used the HMD in our study were color-blind.

4.7 Experimental Procedure

Participants provided informed consent, completed a demographic questionnaire, and were introduced to the robot, including safe kinesthetic guidance in compliant mode. AR headset features were demonstrated for the real-time and combined groups. Participants then completed a brief familiarization session.

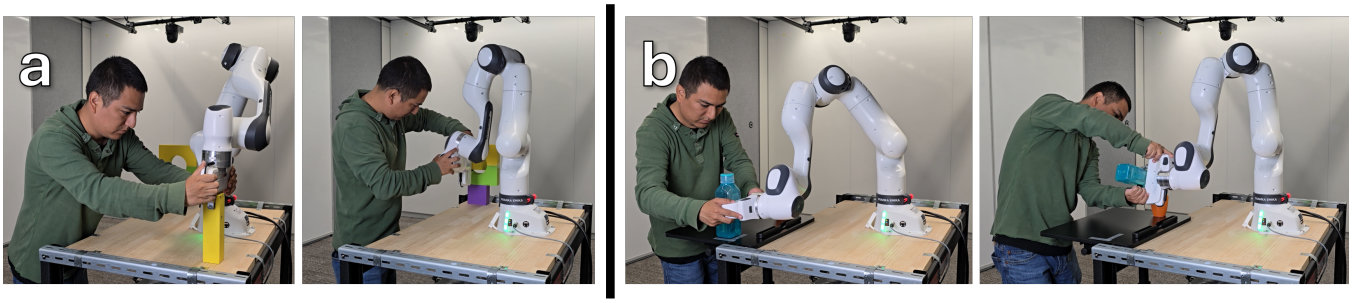


Figure 7: Lab study evaluation tasks: (a) cuboid block insertion and (b) drink pouring.

Tasks were performed sequentially, starting with the easier insertion task followed by pouring, to allow participants to first become familiar with the robot before attempting the more challenging task. Participants were randomly assigned to one of three feedback conditions (offline, real-time, or combined; 12 per group) and instructed to provide six successful demonstrations per task. After the insertion task, SUS and NASA-TLX questionnaires were completed, with NASA-TLX repeated after the pouring task. A brief debriefing interview concluded the one-hour procedure.

For offline participants, feedback consisted of robot execution playback between demonstration rounds. Real-time participants received AR visual cues during demonstrations but no playback. Combined participants received both AR cues and playback. During the pouring task, no feedback was provided for any group, enabling assessment of transfer learning and participants' ability to generalize learned constraints independently.

4.8 Measures

To quantify how each experimental condition affects both human performance and the resulting robot performance, we define a set of objective metrics based on the robot's kinematic state. The mathematical formulation of the metrics is defined in Appendix A.1. In addition, to assess how participants perceived the experience of teaching the robot, we used standard subjective questionnaires.

Task Completion Rate — defined as the proportion of successful demonstrations (i.e., achieving the task objective) to the total number of demonstrations.

Active Demonstration Time (s) — To quantify the effective duration of human demonstrations, we computed the *active demonstration time* by excluding periods in which the robot remained idle. The total idle duration was subtracted from the total demonstration duration to yield the active demonstration time.

Distance to Joint Limits (rad) — We quantified how closely each joint approached its mechanical limits using a normalized distance-to-limit measure, reflecting participants' ability to maintain safe joint mobility within the robot's kinematic range.

Distance to Self-Collision (m) — To assess safe motion behavior, we analyzed the minimum distance between robot links during demonstrations, indicating how conservatively participants configured the robot's posture. Using the MoveIt 2 collision environment[4], each robot state was inserted into the planning scene, and signed distances between all link pairs were computed. Larger values reflect safer kinematic configurations. Analysis focused on the minimum distance across all link pairs for each configuration.

Quality of Demonstration (Manipulability) — To evaluate the dexterity of demonstrated configurations, we used the manipulability measure [90], which reflects proximity to kinematic singularities and has been linked to demonstration quality in learning-from-demonstration settings [13, 62]. Higher manipulability values indicate more dexterous and higher-quality demonstrations, whereas lower values reflect reduced motion flexibility.

Robot Task Completion — We evaluated the robot's ability to execute tasks using models trained (following a method similar to previous works [13, 65]) on each participant's successful demonstrations, excluding failed or excessively long demonstrations. Trajectories were executed on the Franka robot using the ROS2 joint trajectory controller, smoothed with a moving average filter, and temporally aligned via dynamic time warping [70]. For the insertion task, a trial was successful if the robot picked up and partially inserted the block without collisions. For pouring, a trial was successful if the robot grasped and transported the bottle with correct inclination toward the cup, avoiding self-collisions or contact with the setup.

Quality of Task Execution — The manipulability measure was also used to assess the quality of robot task execution, capturing the robot's dexterity and ability to operate away from singular configurations. The policy learning method is detailed in Appendix A.2.

System Usability Scale (SUS) — To evaluate system usability in each condition, we used the SUS [16]. The SUS consists of ten items rated on a 5-point Likert scale and produces a single overall usability score. Its strength lies in being a widely adopted, benchmarked instrument across diverse interface types and applications, enabling us to interpret the usability of our system relative to established norms beyond AR or robotics [47].

Task Load — We employed the NASA-TLX [35] to measure participants' subjective workload across feedback conditions. The NASA-TLX consists of six sub-scales, each rated on a 21-point scale.

In addition to these quantitative metrics, we asked participants about their experience at the end of the experiment in semi-structured interviews (see questions in Appendix A.3).

[4]https://moveit.ai/

4.9 Statistical Analysis

We used linear mixed-effects models to analyze the effects of feedback condition and demonstration number on outcome measures:

$$y \sim \text{condition} + \text{demo\_index} + \text{condition} : \text{demo\_index} + (\text{demo\_index} \mid \text{participant}),$$

where y is the dependent variable, *condition* the experimental condition, and *demo_index* the demonstration number. Random slopes for demonstration number accounted for individual performance trajectories. Post-hoc comparisons used estimated marginal means with Tukey adjustments.

For single-observation measures per participant (e.g., SUS, NASA-TLX, task execution), Welch's ANOVA tested group differences, with Games-Howell post-hoc contrasts adjusting for unequal variances and sample sizes.

4.10 Qualitative Analysis

Audio recordings were transcribed using Otter.ai and reviewed by the first author, with personally identifiable information replaced by pseudonyms (e.g., P29, offline). Two coders independently familiarized themselves with 25% of the transcripts, generating preliminary inductive codes [77]. Codes and relevant quotes were imported into a shared Miro board, and the coders collaboratively developed a unified code set. Remaining transcripts were jointly coded, refining the codes and consolidating themes. All iterations were documented for transparent version control. Candidate themes were then generated, refined, and finalized through discussion, with the second author verifying reported quotes against the original audio before raw data deletion.

5 Results

5.1 Task Completion Rate

Descriptive statistics indicated clear differences in completion rates across conditions and tasks. For the insertion task, completion rates increased from offline ($M = 0.78, SD = 0.14$) to real-time ($M = 0.86, SD = 0.13$) and combined ($M = 0.94, SD = 0.09$). A similar pattern was observed for the pouring task, where the offline condition showed substantially lower performance ($M = 0.49, SD = 0.10$) compared to real-time ($M = 0.83, SD = 0.14$) and combined ($M = 0.82, SD = 0.12$), which were comparable. Welch's ANOVA confirmed a significant effect of group for insertion ($F(2, 20.90) = 5.72, p = 0.010$), with Games-Howell post-hoc tests showing that combined significantly outperformed offline ($p = 0.010$), while other comparisons were not significant. For pouring, a strong group effect was observed ($F(2, 21.47) = 33.53, p < 0.001$), with both combined and real-time significantly outperforming offline (both $p < 0.001$), and no significant difference between combined and real-time ($p = 0.97$).

5.2 Active Demonstration Time

To investigate how the feedback conditions influenced demonstration time, we first analyzed all six successful demonstrations for each task across all conditions. We then conducted separate analyses on the initial and final three demonstrations to examine changes over time. All results reported here focus on active demonstration time (i.e., the periods during which participants were actively guiding the robot). Figure 8 shows the mean active demonstration times for both tasks.

Insertion Task — A mixed-effects analysis of active demonstration time revealed a significant main effect of demonstration number ($\chi^2(1) = 6.39, p = 0.01$), indicating changes as participants progressed through repetitions. No significant effects of condition ($\chi^2(2) = 0.56, p = 0.75$) or condition-by-demonstration interaction ($\chi^2(2) = 0.37, p = 0.83$) were observed. Estimated marginal means showed comparable active times across conditions (offline: $M = 16.7$, real-time: $M = 17.5$, combined: $M = 15.8$; all $p > 0.73$). Analyses of the initial three demonstrations showed a significant effect of demonstration number ($\chi^2(1) = 11.1, p < 0.001$), while no significant effects were found in the final three demonstrations (all $p > 0.15$), suggesting early learning effects that stabilized over time.

Pouring Task — For the pouring task, significant main effects of condition ($\chi^2(2) = 12.43, p = 0.002$) and demonstration number ($\chi^2(1) = 19.02, p < 0.001$) were observed, with no significant interaction ($\chi^2(2) = 0.15, p = 0.92$). Estimated marginal means indicated longer active times in the offline condition ($M = 20.2$) compared to both real-time ($M = 11.8$) and combined ($M = 11.5$) conditions. Tukey-adjusted comparisons confirmed significant differences between offline and both feedback conditions (both $p < 0.001$), with no difference between real-time and combined ($p = 0.97$). These results indicate robust condition-dependent differences alongside a general repetition-related effect. The results are summarized in Table 2.

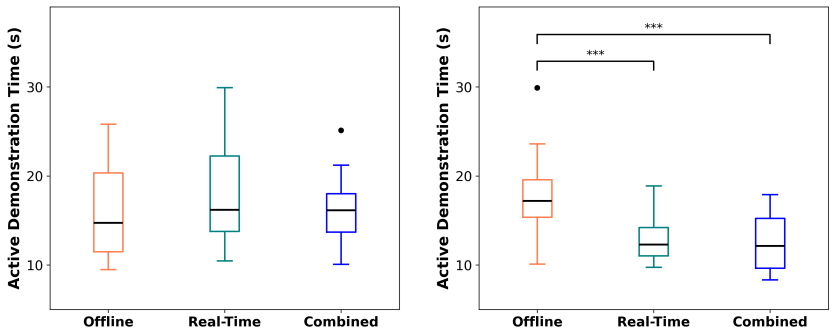
5.3 Distance to Joint Limits

Insertion Task — The analysis revealed significant main effects of condition ($\chi^2(2) = 8.15, p = 0.01$) and demonstration number ($\chi^2(1) = 3.88, p = 0.04$), with no significant condition-by-demonstration interaction ($\chi^2(2) = 1.67, p = 0.43$). Estimated marginal means showed smaller joint-limit clearance in the offline condition ($M = 0.331$) compared to the real-time ($M = 0.341$) and combined ($M = 0.349$) conditions, with a significant difference between offline and combined ($p = 0.01$). Figure 9a shows the normalized distance to joint limits for each condition. Analysis of the initial three demonstrations showed a significant effect of condition ($\chi^2(2) = 9.17, p = 0.01$), whereas no significant effects were observed in the final three demonstrations (all $p > 0.39$), indicating that early group differences in joint-limit clearance diminished over repetitions. The results are summarized in Table 3.

Pouring Task — For the pouring task, no significant main effects of condition ($\chi^2(2) = 1.17, p = 0.55$) or demonstration number ($\chi^2(1) = 2.25, p = 0.13$) were observed, and the interaction was also non-significant ($\chi^2(2) = 1.21, p = 0.54$). Estimated marginal means indicated comparable joint-limit distances across offline ($M = 0.345$), real-time ($M = 0.340$), and combined ($M = 0.334$) conditions, with no significant pairwise differences. These results indicate consistent joint-limit clearance across conditions throughout the pouring task. The results are summarized in Table 3, and Figure 9b presents the normalized distance to joint limits for each condition.

Table 2: Between Conditions analysis results of the linear mixed-effects (LME) model, estimated marginal means (EMMs), and Tukey-adjusted pairwise contrasts for demonstration time (seconds).

		Insertion Task			Pouring Task
		Initial (three demos)	*Last (three demos)*	*All (six demos)*	*All (six demos)*
LME	*Condition*	1.067 ($p = 0.59$)	3.193 ($p = 0.21$)	0.561 ($p = 0.75$)	**12.43 ($p = 0.002$)****
	Demonstration	**11.1 ($p < 0.001$)*****	0.639 ($p = 0.42$)	**6.39 ($p = 0.01$)***	**19.02 ($p < 0.001$)*****
	Condition : Demonstration	1.072 ($p = 0.58$)	3.822 ($p = 0.14$)	0.372 ($p = 0.83$)	0.157 ($p = 0.93$)
EMMs	*Offline Feedback*	17.7 [13.7, 21.6]	13.4 [10.2, 16.7]	16.7 [13.5, 19.9]	20.2 [17.8, 22.6]
	Real-Time Feedback	18.9 [15.0, 22.8]	15.6 [12.4, 18.7]	17.5 [14.2, 20.7]	11.8 [9.51, 14.1]
	Combined Feedback	17.0 [13.1, 20.9]	14.5 [11.3, 17.7]	15.8 [12.6, 19.0]	11.5 [9.21, 13.8]
Contrasts	*Offline — Real-Time*	−1.24 ($p = 0.89$)	−2.14 ($p = 0.60$)	−0.74 ($p = 0.94$)	**8.35 ($p < 0.001$)*****
	Offline — Combined	0.656 ($p = 0.97$)	−1.03 ($p = 0.88$)	0.923 ($p = 0.91$)	**8.66 ($p < 0.001$)*****
	Real-Time — Combined	1.889 ($p = 0.77$)	1.11 ($p = 0.87$)	1.664 ($p = 0.74$)	0.309 ($p = 0.98$)



(a) Insertion task **(b) Pouring task**

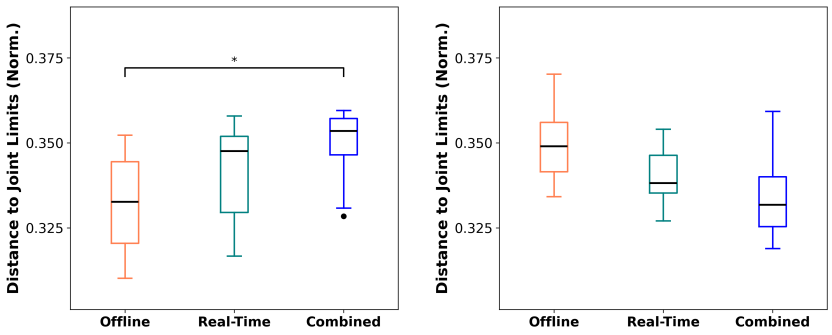
Figure 8: Active demonstration time (seconds) for both tasks: (a) insertion task and (b) pouring task.

Table 3: Between Conditions analysis results of the linear mixed-effects (LME) model, estimated marginal means (EMMs), and Tukey-adjusted pairwise contrasts for distance to joint limits. Values are normalized to the corresponding joint ranges, where smaller values indicate proximity to the joint limits (and larger values indicate being farther away).

		Insertion Task			Pouring Task
		Initial (three demos)	*Last (three demos)*	*All (six demos)*	*All (six demos)*
LME	*Condition*	**9.17 ($p = 0.01$)***	0.396 ($p = 0.82$)	**8.15 ($p = 0.01$)***	1.17 ($p = 0.55$)
	Demonstration	1.230 ($p = 0.26$)	0.392 ($p = 0.53$)	**3.88 ($p = 0.04$)***	2.25 ($p = 0.13$)
	Condition : Demonstration	0.468 ($p = 0.79$)	1.473 ($p = 0.47$)	1.67 ($p = 0.43$)	1.21 ($p = 0.54$)
EMMs	*Offline Feedback*	0.330 [0.322, 0.339]	0.337 [0.330, 0.347]	0.331 [0.323, 0.340]	0.345 [0.338, 0.353]
	Real-Time Feedback	0.340 [0.331, 0.349]	0.342 [0.334, 0.350]	0.341 [0.333, 0.350]	0.340 [0.333, 0.347]
	Combined Feedback	0.349 [0.341, 0.359]	0.351 [0.342, 0.358]	0.349 [0.341, 0.358]	0.334 [0.326, 0.341]
Contrasts	*Offline — Real-Time*	−0.009 ($p = 0.29$)	−0.004 ($p = 0.78$)	−0.009 ($p = 0.24$)	0.005 ($p = 0.57$)
	Offline — Combined	**−0.019 ($p = 0.01$)***	−0.012 ($p = 0.12$)	**−0.018 ($p = 0.01$)***	0.011 ($p = 0.08$)
	Real-Time — Combined	−0.009 ($p = 0.26$)	−0.007 ($p = 0.37$)	−0.008 ($p = 0.35$)	0.006 ($p = 0.42$)

5.4 Distance to Self-Collision

Insertion Task — No significant main effects of condition ($\chi^2(2) = 2.25$, $p = 0.32$) or demonstration number ($\chi^2(1) = 0.77$, $p = 0.37$) were observed, and the condition-by-demonstration interaction was not significant ($\chi^2(2) = 1.46$, $p = 0.48$). Estimated marginal



(a) Insertion task **(b) Pouring task**

Figure 9: Distance to joint limits for both tasks: (a) insertion task and (b) pouring task.

means indicated comparable self-collision distances across conditions, with a small but significant difference between offline and real-time conditions ($p = 0.04$). Analyses of the initial three demonstrations revealed no significant effects (all $p > 0.35$), whereas for the final three demonstrations, the effect of demonstration number approached significance ($\chi^2(1) = 3.81$, $p = 0.05$), and a significant difference emerged between offline and real-time conditions ($p = 0.01$). These results suggest largely stable self-collision margins with a modest late divergence between conditions. Figure 10a shows the distance to self-collision for each condition.

Pouring Task — For the pouring task, a significant main effect of condition was observed ($\chi^2(2) = 7.95$, $p = 0.019$), while the effects of demonstration number ($\chi^2(1) = 1.27$, $p = 0.25$) and the interaction term ($\chi^2(2) = 1.62$, $p = 0.44$) were not significant. Estimated marginal means showed larger self-collision clearance in the offline condition ($M = 0.124$) compared to the real-time condition ($M = 0.107$), with a significant pairwise difference ($p = 0.005$); no other contrasts were significant. These findings indicate condition-dependent differences in self-collision safety margins during the pouring task, with offline demonstrations maintaining the greatest clearance. The results are summarized in Table 4.

5.5 Quality of Demonstration

To assess the quality of the demonstrated configurations, we analyzed manipulability as an indicator of how dexterous and well-conditioned the robot's posture was during demonstrations. Figure 11a shows the manipulability measure of each condition and task.

Insertion Task — No significant main effects of condition ($\chi^2(2) = 1.75$, $p = 0.42$) or demonstration number ($\chi^2(1) = 0.47$, $p = 0.49$) were observed, and the condition-by-demonstration interaction was not significant ($\chi^2(2) = 0.84$, $p = 0.66$). Estimated marginal means were similar across offline ($M = 0.0497$), real-time ($M = 0.0512$), and combined ($M = 0.0493$) conditions, with no significant pairwise differences (all $p > 0.57$). Separate analyses of the initial and final three demonstrations likewise revealed no significant effects or pairwise differences (all $p > 0.11$), indicating stable manipulability across conditions and repetitions.

Pouring Task — For the pouring task, the mixed-effects model revealed no significant main effects of condition ($\chi^2(2) = 0.99$, $p = 0.61$) or demonstration number ($\chi^2(1) = 1.74$, $p = 0.19$), and no significant interaction ($\chi^2(2) = 1.61$, $p = 0.44$). Estimated marginal means were comparable across offline ($M = 0.0684$), real-time ($M = 0.0705$), and combined ($M = 0.0725$) conditions, with no significant pairwise differences. These results indicate consistent robot manipulability across conditions throughout the pouring task.

5.6 Robot Performance

Task performance differed across conditions. For the insertion task, success rates were similar across offline (9/12), real-time (8/12), and combined (9/12) conditions. In contrast, the pouring task showed clearer differences, with higher success under real-time (11/12) and combined (10/12) conditions compared to offline (8/12). Successful autonomous task executions are shown in Figure 12.

5.7 Quality of Task Execution

Figure 11b shows a summary of the manipulability measure across conditions and tasks.

Insertion Task — A Welch's ANOVA found no significant effect of condition on task execution quality ($F(2, 20.47) = 0.332$, $p = 0.72$), with no significant pairwise differences between any conditions (Games-Howell, all $p \geq 0.69$).

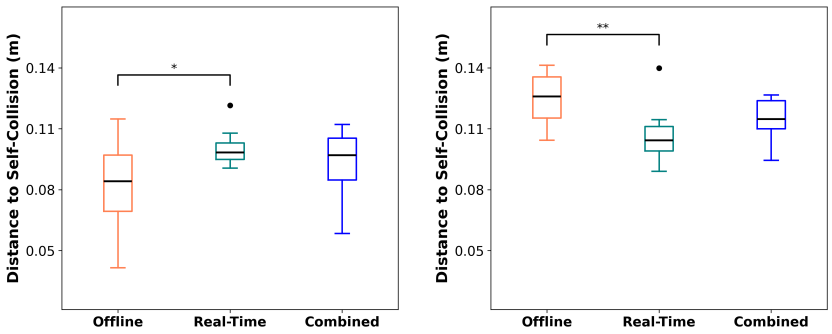
Pouring Task — Likewise, no significant effect of condition was observed ($F(2, 21.62) = 0.510$, $p = 0.51$), and post-hoc comparisons revealed no significant differences between conditions (all $p \geq 0.57$).

5.8 System Usability

A Welch's ANOVA revealed a significant effect of condition on system usability scores, $F(2, 20.35) = 8.71$, $p < 0.001$. The offline condition reported the lowest usability scores ($M = 40.62$, $SD = 22.67$, $95\% CI = [26.22, 55.03]$), followed by the real-time condition ($M = 63.33$, $SD = 14.03$, $95\% CI = [54.42, 72.25]$) and the combined condition ($M = 70.83$, $SD = 10.19$, $95\% CI = [64.36, 77.31]$). Games-Howell post-hoc tests showed that the offline condition scored significantly lower than the real-time condition ($p = 0.02$) and the combined condition ($p = 0.002$). The difference between real-time and combined conditions was not significant ($p = 0.31$). Figure 13 shows a summary of SUS scores across conditions.

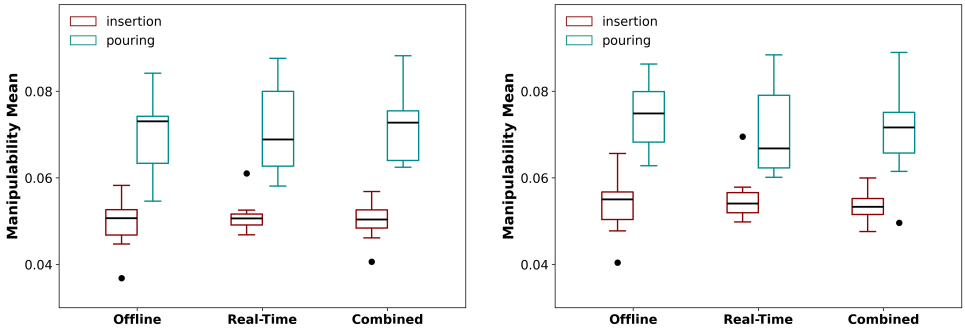
Table 4: Between Conditions analysis results of the linear mixed-effects (LME) model, estimated marginal means (EMMs), and Tukey-adjusted pairwise contrasts for distance to self-collision (m).

		Insertion Task			Pouring Task
		Initial (three demos)	*Last (three demos)*	*All (six demos)*	*All (six demos)*
LME	*Condition*	2.087 ($p = 0.35$)	4.145 ($p = 0.13$)	2.25 ($p = 0.32$)	**7.95 ($p = 0.01$)***
	Demonstration	0.044 ($p = 0.83$)	**3.814 ($p = 0.05$)***	0.77 ($p = 0.37$)	1.27 ($p = 0.25$)
	Condition : Demonstration	0.730 ($p = 0.69$)	0.740 ($p = 0.70$)	1.46 ($p = 0.48$)	1.62 ($p = 0.44$)
EMMs	*Offline Feedback*	0.084 [0.074, 0.095]	0.075 [0.063, 0.087]	0.082 [0.071, 0.092]	0.124 [0.116, 0.131]
	Real-Time Feedback	0.099 [0.089, 0.109]	0.101 [0.089, 0.112]	0.099 [0.089, 0.110]	0.107 [0.099, 0.114]
	Combined Feedback	0.092 [0.082, 0.102]	0.093 [0.081, 0.106]	0.092 [0.082, 0.103]	0.114 [0.107, 0.121]
Contrasts	*Offline — Real-Time*	−0.014 ($p = 0.11$)	**−0.026 ($p = 0.01$)***	**−0.018 ($p = 0.04$)***	**0.017 ($p = 0.005$)****
	Offline — Combined	−0.008 ($p = 0.53$)	−0.018 ($p = 0.09$)	−0.011 ($p = 0.32$)	0.009 ($p = 0.16$)
	Real-Time — Combined	0.007 ($p = 0.61$)	0.007 ($p = 0.64$)	0.008 ($p = 0.57$)	−0.007 ($p = 0.28$)



(a) Insertion task **(b) Pouring task**

Figure 10: Distance to self-collision for both tasks: (a) insertion task and (b) pouring task.



(a) Manip. measure of input demonstrations **(b) Manip. measure of task execution**

Figure 11: Manipulability measure of (a) input demonstrations and (b) robot task execution.

5.9 NASA-TLX

Welch's ANOVAs were conducted to examine differences across the three feedback conditions for all NASA-TLX sub-scales in the *insertion* and *pouring* tasks. For the insertion task, no significant condition effects were observed on any TLX dimension (all $p > 0.05$). In contrast, for the pouring task, significant condition effects were found for Mental Demand ($F(2, 21.83) = 6.69$, $p = 0.005$), Physical Demand ($F(2, 21.68) = 3.48$, $p = 0.049$), Performance ($F(2, 21.11) = 7.98$, $p = 0.003$), Effort ($F(2, 21.71) = 8.03$, $p = 0.002$), and Frustration ($F(2, 21.98) = 10.50$, $p < 0.001$), while Temporal Demand was not significant ($p = 0.17$).

Post-hoc Games-Howell tests for the pouring task showed that the offline condition reported significantly higher workload than both real-time and combined conditions across Mental Demand,

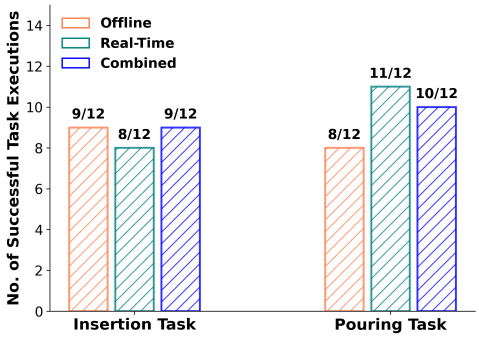


Figure 12: Robot task executions.

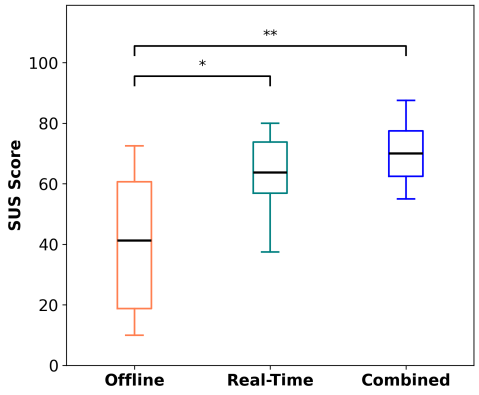


Figure 13: System Usability Score (SUS).

Performance, Effort, and Frustration (all $p \leq 0.021$), and higher Physical Demand than the real-time condition ($p = 0.043$). No significant post-hoc differences were found for the insertion task. Overall, these results indicate pronounced workload differences during the pouring task, with the offline condition consistently associated with higher perceived workload, whereas no meaningful between-condition differences emerged for the insertion task (see Table A1).

5.10 Qualitative Results

5.10.1 Adjustments After Failed Demonstrations.

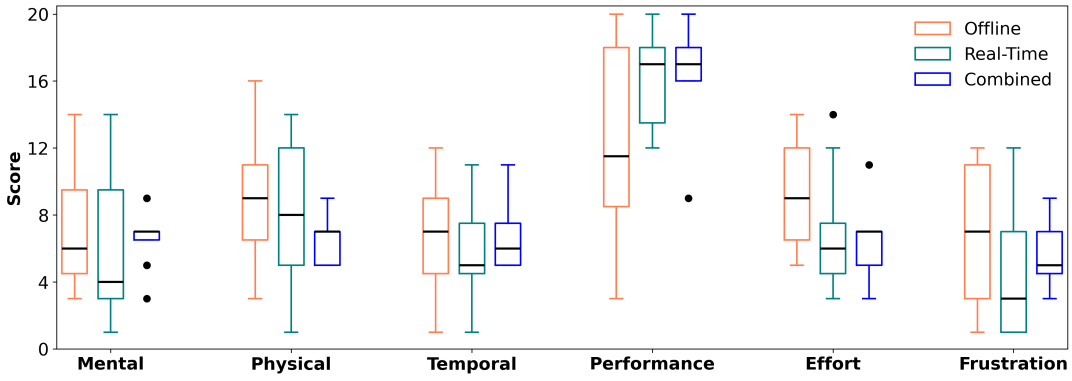
Proactive Avoidance of Joint Limits. Participants adjusted their demonstrations to avoid joint-limit violations, with awareness shaped by feedback modality. In the offline condition, they learned constraints retrospectively through failure and playback: "*tried to move the arm carefully not to go beyond its joint's limit ... because that was the problem*" (P01, offline); "*learned about the joint limits ... and tried to make moves where I have more space to move*" (P05, offline). With real-time feedback, participants relied on AR cues to guide motion: "*following the limit given by every joint's information ... basically I just follow anything that AR information gives to me*" (P17, real-time); "*find an approach where the robot does not reach its limitations*" (P14, real-time). Combined feedback supported integrated understanding, enabling intentional adaptation: "*I learned about each joint's movement ... and then I adjust my actions for the robot to move as efficient and effective as possible*" (P27, combined).

Re-sequencing and Decomposition of Movements. After failed demonstrations, participants adjusted joint movement order, with strategies varying by feedback type. Offline participants made broad sequencing changes: "*completely changing the starting position ... and making it go in other direction*" (P06, offline); "*adjust[ing] the link joints before adjusting the gripper*" (P04, offline). Real-time feedback led to more structured strategies guided by AR cues: "*rotate the bigger ... lower joint first and then go up ... the bottom one is more certain*" (P16, real-time); "*there is a limit for upward movement, so I have to move it downward and twist it*" (P18, real-time). Combined feedback supported fine-grained decomposition: "*change the sequence and control the joint ... only changed one joint at one time*" (P29, combined).

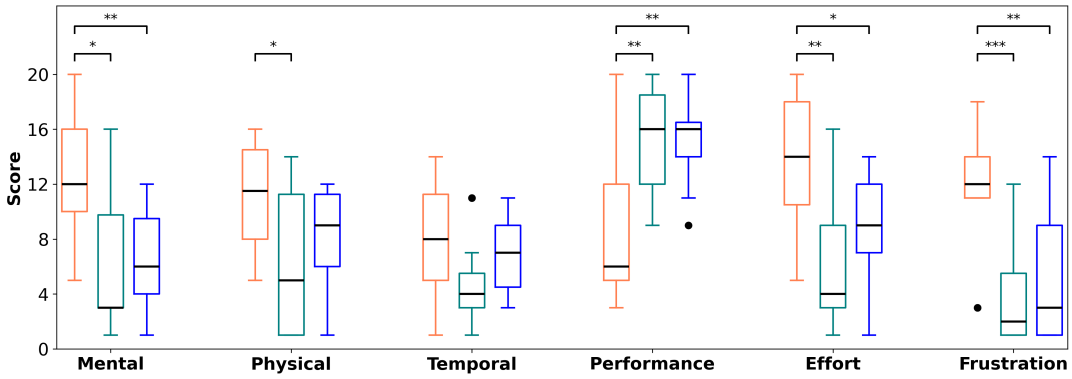
Conservative and Simplified Motion Strategies. After failing to record a successful demonstration, participants frequently adopted more conservative control strategies, including slowing down and simplifying movements. Participants in the offline condition emphasized general caution, stating they would be "*more cautious of the limits ... and keep the movements fluid*" (P3, offline) or "*try to make single directions compared to diagonal ones*" (P8, offline). Participants in the real-time feedback condition linked conservative behavior to specific failure events detected through AR. One participant recalled a failure due to "*a collision of the block with the table*" and another due to "*rushing too hard to get the block*" (P15, real-time), prompting more controlled subsequent attempts. In the combined condition, slowing down was sometimes associated with increased cognitive demand, as one participant noted they "*tried to do it slowly*" because the feedback "*was kind of making me a little bit miserable*" (P35, combined).

5.10.2 Understanding Key Qualities of an Effective Demonstration. Across all conditions, participants showed understanding of the robot's physical constraints, particularly joint limits and ranges of motion, as the most fundamental quality of an effective demonstration. In the offline condition, this understanding emerged primarily through reflection on failures and robot rollouts. Participants emphasized avoiding problematic configurations, noting that "*the joints were the main problem ... I had to move it in such a way that it did not go beyond its limit*" (P1, offline) and recognizing that "*there is some angle that it cannot be worked in that way*" (P7, offline). Participants in both real-time and combined feedback conditions described AR as making these constraints immediately visible and actionable. One participant explained that seeing "*the movement limit of every joint ... helps me a lot to determine the extent ... I should move every joint*" (P17, real-time).

Learning and Internalization of Robot Capabilities. Participants in conditions with real-time AR feedback more explicitly described learning about the robot's capabilities rather than simply correcting errors. They reflected on gaining familiarity with the joints' behavior over time, noting that after initial failures they "*already know the limitation of the robot itself and know how to move the robot more efficient[ly]*" (P18, real-time). Participants also described effective demonstrations as those that supported learning through direct,



(a) Insertion task.



(b) Pouring task.

Figure 14: Participants' responses to the NASA–TLX questionnaire (21-point scale).

embodied interaction with the robot, rather than abstract or theoretical instruction. One participant emphasized that understanding emerged through experience: "*getting a hands-on ... experiencing it myself ... if I was told ... theoretically, I don't think I would understand it*" (P14, real-time). Others noted that real-time feedback supported early familiarization and confidence building, with one participant reporting greater confidence during tasks performed with AR compared to those without it.

5.10.3 *Perceived Impact of Demonstrations on Robot Performance.*

Demonstration Quality Directly Shapes Robot Effectiveness and Efficiency. Participants commonly perceived a strong link between the quality of their demonstrations and the robot's resulting performance, particularly in terms of efficiency, smoothness, and task success. Participants across conditions emphasized that demonstrations did not merely teach the robot what to do, but how well to do it. One participant reflected that effective teaching involved ensuring the robot learned "*the most efficient movement as possible ... not only just to solve the challenge, but ... in the most perfect way*" (P27, combined). Similarly, others noted that demonstrations became more refined over time after the second round, with one participant stating that "*the robot learned more with more trials*" (P8, offline). Several participants associated improved robot performance with clearer, more confident demonstrations, suggesting that inefficiencies or uncertainty during demonstration could negatively affect task execution: "*if there is some uncertainty ... then the robot does not have enough ability to complete the task*" (P2, offline).

Learning as an Iterative Process Accumulating Across Demonstrations. Participants frequently described the robot's learning as incremental, improving as more demonstrations were provided and as their own performance stabilized. Many participants believed that repetition and refinement across trials were key to improving the robot's performance. One participant explained that as their own performance improved, "*the speed increased, the efficiency increased ... if the robot considers the last few attempts ... it would perform very well*" (P19, real-time). Another similarly noted that with "*more iterations ... it will know how it will be done easily*" (P26, combined). This perception was shared across conditions, with participants suggesting that later demonstrations carried more useful information than early exploratory ones. However, some participants also expressed uncertainty about how demonstrations were weighted internally, particularly when early and late demonstrations differed substantially.

Partial Understanding and Uncertainty About Robot Interpretation. Despite generally positive perceptions, participants often expressed

uncertainty about how accurately or consistently the robot interpreted their demonstrations. Some participants questioned whether the robot truly "learned" from them or merely recorded motions. One participant remarked, "*I'm not sure if it's learning from me ... it's definitely recording what I'm doing*" (P14, real-time), suggesting ambiguity in the learning process. Others noticed discrepancies between their demonstrations and the robot execution, observing that it does not "*mimic your movement exactly ... there is some noise or some extra movement*" (P22, real-time). Participants in the combined condition echoed this mixed understanding, noting that while the robot was generally responsive, there was "*a little bit of ... inaccuracy towards the end*" (P34, combined). Several participants attributed these issues to inconsistent demonstrations or insufficient system understanding of task context, rather than to outright failure.

5.10.4 Role of Feedback in Supporting Performance on the Second Task.

Internalization of Joint Limits and Range of Motion. Across all conditions, participants reported that feedback (real-time and/or offline) during the insertion task helped them internalize the robot's joint limits and range of motion, which directly supported more effective performance in the second task. Participants in the offline condition described gaining an understanding of the robot's movement capabilities through experience and repetition, enabling smoother performance later. One participant noted that they "*knew the range of motion of the robot ... how should I move it first so that it can do the second motion ... effectively*" (P6, offline), while another stated they could "*navigate it relatively easier than the first two tasks*" after understanding joint behavior (P3, offline). This effect was especially pronounced in the real-time feedback condition, where AR made joint limits explicit. Participants frequently reported remembering these limits even after the AR was removed, stating that "*by the time I removed the AR headset ... I already know the limitation of the robot*" (P20, real-time) and that AR helped them "*try to not ... reach the limit*" (P23, real-time) during the second task. Similarly, participants in the combined condition emphasized that AR allowed them to "*see the approximate limits ... so in the second task ... I don't push it to its limits*" (P36, combined).

AR Feedback as a Scaffold for Learning and Confidence Building. Participants widely perceived AR feedback as a learning scaffold that supported confidence. Visualization participants described AR as intuitive and informative, noting that it allowed them to "*visually see how much I was moving ... and understand to what level I can go*" (P15, real-time). Several participants emphasized that this scaffolding was especially valuable for the more complex second task, which involved coordinated rotations and precise manipulation: "*for the second task ... it's more complicated ... you have to rotate ... and then pour*" (P16, real-time). However, participants also acknowledged limitations of relying solely on AR. One participant (P17, real-time) estimated that AR provided "*70% of the information,*" but without continuous feedback they had to rely on memory and prediction, which "*becomes a bit complicated*". This suggests that AR supported learning and confidence, but its removal increased cognitive demands during task execution.

6 Discussion

This work set out to explore how robots can communicate their internal constraints to novice users during kinesthetic Learning from Demonstration, and how such communication reshapes the interaction between human teachers and robot learners. Rather than treating feedback as a post-hoc diagnostic tool or a purely technical aid, our findings position real-time AR-based feedback as an interaction design strategy for making robot learning legible during embodied teaching. In this section, we discuss how our results contribute design knowledge by contextualizing constraint feedback as a form of dialogue, examining how actionable and escalating cues shape user behavior, and articulating transferable design principles for communicating invisible system states.

6.1 Effects of Constraint Visualizations

Joint Limits Visualization. The joint-limit visualization showed a significant impact on user interaction, with participants consistently adjusting their demonstrations to avoid exceeding joint limits. In the real-time and combined conditions, users relied on AR cues to proactively guide motion, whereas offline participants developed this understanding retrospectively through failures and playback (see Section 5.3 for quantitative results). Participants also reported internalizing joint limits over time and transferring this knowledge to subsequent tasks (see Section 5.10.1 for qualitative results), indicating that this visualization strongly supported both immediate performance and learning.

Self-Collision Visualization. The self-collision visualization showed a more nuanced impact on user interaction, with evidence emerging across both quantitative and qualitative findings. Quantitatively, improvements in performance were less pronounced compared to the joint-limit visualization, suggesting a more limited effect on proactive behavior (see Section 5.4). Qualitatively, however, participants demonstrated increased awareness of collisions through reflections on failure events (see Section 5.10.2). In the real-time condition, users explicitly linked failures to their ways of demonstrating the given tasks and subsequently adopted more cautious and controlled strategies, such as slowing down or simplifying movements. This indicates that while the visualization supported users in recognizing and responding to collision-related issues, its influence was primarily reactive, with less evidence of proactive planning or anticipation.

Manipulability Visualization. The manipulability visualization did not show a significant impact on user performance. This is likely due to the nature of the task, which inherently constrained users from entering near-singular configurations (i.e., extreme cases where the end effector's movement is reduced), thereby limiting opportunities for the visualization to influence behavior. Quantitatively, no significant differences were observed across conditions (see Section 5.5), and qualitatively, participants did not explicitly report awareness of manipulability-related concepts. Future work should consider designing experiments in which manipulability is investigated independently, allowing for a more focused and rigorous evaluation of its impact on user understanding and performance.

6.2 Designing for Mutual Legibility in Kinesthetic Teaching

A central challenge in kinesthetic robot teaching is that users must reason about constraints (e.g., joint limits, self-collisions, singularities) that are not directly perceptible through physical interaction alone. While users can feel resistance or abrupt stopping, such sensations provide little insight into why a configuration is problematic or how it might be corrected. Our findings suggest that these breakdowns are less about user skill and more about a lack of mutual legibility between human and robot. By visualizing internal robot states in situ, the AR-based feedback system transformed teaching from a one-sided act of demonstration into a more reciprocal interaction. Instead of guessing what the robot could or could not accommodate, users were able to see how their actions related to the robot's internal constraints. This aligns with previous works in HRI on intelligibility and explainable interaction: the robot does not merely execute motions but actively "shows" its limits [10], enabling users to adjust their behavior accordingly.

Importantly, this shift reframes **Learning from Demonstration as a communicative process rather than a mere data collection step**. Teaching quality improved not only because users avoided errors, but because the interaction itself supported sensemaking about how the robot experiences the task.

6.3 From Awareness to Action: Designing Actionable Constraint Feedback

The focus group study revealed that simply making constraints visible is insufficient. Participants consistently emphasized that awareness cues, such as color changes indicating proximity to a limit, must be coupled with guidance that helps users recover from problematic configurations. This distinction between knowing that something is wrong and knowing what to do next proved critical in the lab study. Our AR system operationalized this insight by layering feedback: initial awareness was communicated through subtle, color-coded cues, while more explicit guidance (e.g., arrows or intensified highlights) emerged as users approached or crossed critical thresholds. The study results suggest that this approach supported smoother demonstrations and reduced time spent in unproductive or constrained configurations.

From an interaction design perspective, this highlights the importance of **feedback that is not merely informative but actionable**. Effective real-time feedback must support decision-making in the moment, translating abstract constraints into concrete possibilities for movement. This principle extends beyond robotics to other interactive systems where users must navigate invisible or complex limitations in real-time [41, 71].

6.4 Escalation, Minimalism, and Managing Cognitive Load

A recurring concern in both the focus group and prior XR literature is the risk of overwhelming users with too much visual information. Our design addressed this tension through escalation and progressive disclosure: feedback remained invisible during safe operation and emerged gradually as constraints became relevant. This strategy proved effective in balancing awareness and cognitive load. Participants reported lower perceived workload, and observational data suggested that users were rarely distracted by the AR overlays when the robot operated within safe regions. By anchoring feedback spatially to the robot and limiting its presence to moments of risk, the system avoided becoming a constant source of visual noise. From a design standpoint, this demonstrates how **escalation can function as a general interaction pattern for XR systems** [74]. Rather than presenting all available information upfront, systems can remain visually quiet and intervene only when user action requires guidance [49]. This approach respects user attention while preserving access to critical information when it matters most.

6.5 Feedback as a Scaffold for Learning, Not a Crutch

One of the most revealing aspects of the study was the second task, in which all feedback was removed. Users who had previously received real-time AR-based feedback continued to demonstrate safer and more effective teaching behaviors, suggesting that they had internalized aspects of the robot's constraints. This finding reframes real-time feedback not as a permanent assistive layer, but as a temporary learning scaffold beyond general task practice. By receiving immediate, context-specific feedback during early interactions, users developed mental models of the robot's movements and limitations. Once these models were established, users were able to apply them even without visual support.

This distinction highlights **feedback design as a means of supporting learning and understanding rather than dependency**. This aligns with scaffolding theory, in which instructional support is provided during initial learning and gradually withdrawn as learners internalize task structure [83, 88]. In contrast to offline execution-based feedback, which supports reflection after training, real-time feedback enabled learning through embodied action, aligning with theories of experiential and situated learning [42, 44]. By coupling perception and action in real time, feedback becomes part of the task itself, supporting the development of sensorimotor understanding that is difficult to achieve through post-hoc reflection alone.

Our proposed real-time feedback system is designed to support user understanding of the constraints instead of reducing exploration during kinesthetic demonstrations. Furthermore, rather than constraining the search space, the system provides visual feedback that helps users provide effective demonstrations considering the robot's inherent constraints.

6.6 Design Principles for Communicating Invisible Constraints during LfD for Novice Users

Synthesizing findings from both studies, we adapt Vi et al.'s 11 guidelines for XR applications on head-mounted displays [81] to ***LfD tasks that involve novice users*** and propose five guidelines.

- **Design flexible spatial environments to maximize efficiency.** In XR systems for LfD, users teach robots by demonstrating task-specific actions. Efficiency in this context can be measured through the number of tasks completed over

time (task completion rate) and time required to complete a single demonstration (active demonstration time).
Our findings show that real-time and combined feedback conditions significantly outperform offline approaches (see Section 5.1 and 5.2). Both real-time and combined setups lead to higher task completion rates and reduced demonstration time, with no significant performance difference between them.
From a spatial design perspective, these results suggest that real-time feedback alone is sufficient to maximize efficiency. Incorporating mixed elements, such as switching between XR and external displays for task execution verification, does not provide additional efficiency benefits, while introducing unnecessary spatial and cognitive overhead.

- **Use cues to help through experiences which build upon real-world knowledge.** In XR systems for LfD, cues should align with users' real-world knowledge to reduce cognitive load and improve understanding. Our findings show that **color-based indicators** (e.g., green-to-red) are effective for communicating joint limits, **overlay visualizations** help identify self-collisions directly on the robot, and **ellipsoids at the end-effector** intuitively represent manipulability.
- **Balance Comfort and Cognitive Load Through Minimal and Contextual Feedback.** Kinesthetic teaching in LfD is physically and cognitively demanding. XR feedback should minimize strain by avoiding excessive overlays, respecting personal space, and reducing unnecessary movement. Using progressive disclosure, feedback should remain unobtrusive during safe operation and appear only when relevant, ensuring both physical comfort and manageable cognitive load.
- **Design Feedback that is Accurate, Consistent, and Reflective of Robot Capabilities.** XR feedback must faithfully represent the robot's kinematics and limitations to build trust and correct understanding. Providing consistent, real-time responses with clear state distinctions (e.g., safe, near-limit, violation) helps users develop reliable mental models of how the robot behaves and what actions are feasible.
- **Support Learning Through Scaffolded Feedback and Exploration.** LfD is fundamentally a learning process. XR feedback should act as a scaffold by providing strong, real-time guidance during early interactions and gradually reducing support as users gain experience. This enables users to internalize robot constraints and continue effective teaching even without feedback.

While derived from a robotics context, these principles can be applied broadly to interactive systems that need to externalize hidden system states. For example, through XR interfaces that visually embed system intent and status in the user's environment, improving situational awareness [75, 86]; auditory sonification that conveys latent states or system uncertainty via sound [9]; or motion and expressive behaviors that make internal processes interpretable through movement patterns [48].

6.7 Comparison with Existing Visualization Tools

Our work specifically targets the visualization of robot-inherent constraints—such as joint limits and self-collisions—during kinesthetic demonstrations, which is not the primary focus of existing tools like MoveIt. While MoveIt provides valuable visualizations of kinematic structures (e.g., frames and link relationships), it does not explicitly convey the current state of joint angles in a way that is directly interpretable during interaction.

In contrast, our approach introduces explicit, real-time visual encodings of joint angle limits, enabling users—particularly novices—to better understand how close they are to constraint boundaries and thus perform more dexterous and controlled demonstrations. Although prior tools provide baseline visualizations, they are not tailored to this interaction context or user group. Our design complements existing systems and can be integrated into MoveIt to enhance understanding of robot constraints.

6.8 Limitations and Future Work

While the results demonstrate the potential of our AR-based approach, they also present several limitations related to the study population, experimental design, and technical constraints of the AR system, and highlight directions for future work.

- **Generalizability and Study Design.** This study involved university students, which may limit generalizability to professional settings. Future work will evaluate industrial and healthcare workers to assess domain-specific effects. The manipulability visualization introduces an abstract representation that may be interpreted differently by novice users, potentially contributing to variability in user performance. Although notable differences in task completion times were observed, more detailed statistical analysis of this variability remains for future work. We will also incorporate established guidelines from information visualization and accessibility to further improve interface design and usability [54, 87].
- **AR System and Tracking Limitations.** We used state-of-the-art hardware; however, occasional instability in the Meta Quest 3 affected AR consistency. Our experimental setup relied on a single QR-code-based spatial anchor for AR alignment, which exhibited occasional alignment drift (< 1 cm), particularly when out of view, potentially impacting the perception of spatial constraints. To mitigate this, we implemented continuous QR tracking and designed the AR-based task such that the QR code remained within the user's field of view for most of the interaction. Despite these measures, residual instability may still have influenced user performance, highlighting device reliability as a critical factor in AR-based experiments.
- **Ergonomics and User Comfort.** The headset's weight (515 g) may cause discomfort and motion-sickness symptoms such as dizziness and headaches. While this was minimized in our study due to short, simple tasks (approximately 30 minutes), longer or more demanding use may increase both physical discomfort and simulator sickness.

7 Conclusion

This work examined how real-time feedback can function as a communicative and educational mechanism in embodied human-agent/robot interaction. By designing and evaluating an AR-based system that makes otherwise invisible robot constraints perceptible during kinesthetic teaching, we showed how feedback can support safer, more efficient demonstrations while helping users form more accurate mental models of the robot's capabilities.

Our contribution lies in reframing constraint feedback as an interaction design problem, empirically grounding design principles for real-time and escalating feedback, and demonstrating how such feedback can scaffold user learning rather than foster long-term dependence. While evaluated in a robotic teaching context, these insights extend to a broader class of interactive systems in which users must engage with complex, hidden system states through repeated interactions.

Several directions for future research emerge from this work. First, longitudinal studies could examine how real-time feedback shapes learning and skill retention over extended use, including whether and when such feedback should be gradually withdrawn or adapted. Second, future systems could explore adaptive or personalized feedback strategies that respond to user expertise, task complexity [82], or contextual risk [43]. Finally, extending these design principles beyond robotics to domains such as physical rehabilitation or human-AI/robot collaboration [58, 60], offers opportunities to further investigate how real-time, perceptual feedback can support understanding, trust, and agency in interactions between humans and autonomous systems.

Acknowledgments

This work was partially supported by the Australian Research Council (Grants No.: DE210100858; CE260100108; DP260101082; FT250100459).

References

[1] Daron Acemoglu and Pascual Restrepo. 2020. Robots and Jobs: Evidence from US Labor Markets. *Journal of Political Economy* 128, 6 (2020), 2188–2244. doi:10.1086/705716

[2] Baris Akgun, Maya Cakmak, Jae Wook Yoo, and Andrea Lockerd Thomaz. 2012. 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)*. 391–398. doi:10.1145/2157689.2157815

[3] Jeffrey Allen, James E. Young, Daisuke Sakamoto, and Takeo Igarashi. 2012. Style by Demonstration for Interactive Robot Motion. In *Proceedings of the Designing Interactive Systems Conference (DIS '12)*. 592–601. doi:10.1145/2317956.2318045

[4] Patrícia Alves-Oliveira, Maria Luce Lupetti, Michal Luria, Diana Löffler, Mafalda Gamboa, Lea Albaugh, Waki Kamino, Anastasia K. Ostrowski, David Puljiz, Pedro Reynolds-Cuéllar, Marcus Scheunemann, Michael Suguitan, and Dan Lockton. 2021. Collection of Metaphors for Human-Robot Interaction. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference (DIS '21)*. 1366–1379. doi:10.1145/3461778.3462060

[5] Sushilkumar Ambhore. 2020. A Comprehensive Study on Robot Learning from Demonstration. In *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*. IEEE, 291–299. doi:10.1109/ICIMIA48430.2020.9074946

[6] Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35, 4 (2014), 105–120. doi:10.1609/aimag.v35i4.2513

[7] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. 1–13. doi:10.1145/3290605.3300233

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

[9] Simone Arreghini, Antonio Paolillo, Gabriele Abbate, and Alessandro Giusti. 2024. Hearing the Robot's Mind: Sonification for Explicit Feedback in Human-Robot Interaction. In *International Workshop on Human-Friendly Robotics*. Springer, 45–57. doi:10.1007/978-3-031-81688-8_4

[10] Anna Belardinelli, Chao Wang, Daniel Tanneberg, Stephan Hasler, and Michael Gienger. 2025. Train your robot in AR: insights and challenges for humans and robots in continual teaching and learning. *Frontiers in Robotics and AI* 12 (2025). doi:10.3389/frobt.2025.1605652

[11] Jacqueline A. Belzile and Gunilla Öberg. 2012. Where to begin? Grappling with how to use participant interaction in focus group design. *Qualitative Research* 12, 4 (2012), 459–472. doi:10.1177/1468794111433089

[12] Muhammad Bilal, D. Antony Chacon, Nir Lipovetzky, Denny Oetomo, and Wafa Johal. 2026. Investigating the Impact of Robot Degree of Redundancy on Learning from Demonstration. In *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI '26)*. 825–833. doi:10.1145/3757279.3785606

[13] Muhammad Bilal, Nir Lipovetzky, Denny Oetomo, and Wafa Johal. 2024. Beyond Success: Quantifying Demonstration Quality in Learning from Demonstration. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 5120–5127. doi:10.1109/IROS58592.2024.10802187

[14] Aude Billard, Sylvain Calinon, Ruediger Dillmann, and Stefan Schaal. 2008. Survey: Robot Programming by Demonstration. *Springer Handbook of Robotics* (2008), 1371–1394. doi:10.1007/978-3-540-30301-5_60

[15] Erik A. Billing and Thomas Hellström. 2010. A Formalism for Learning from Demonstration. *Paladyn, Journal of Behavioral Robotics* 1, 1 (2010), 1–13. doi:10.2478/s13230-010-0001-5

[16] John Brooke. 1996. SUS: A 'Quick and Dirty' Usability Scale. *Usability Evaluation in Industry* 189, 194 (1996), 4–7. doi:10.1201/9781498710411-35

[17] Sylvain Calinon. 2016. A tutorial on task-parameterized movement learning and retrieval. *Intelligent Service Robotics* 9 (2016), 1–29. doi:10.1007/s11370-015-0187-9

[18] Sylvain Calinon. 2018. Learning from Demonstration (Programming by Demonstration). *Encyclopedia of Robotics* (2018), 1–8. doi:10.1007/978-3-642-41610-1_27-1

[19] Sylvain Calinon, Florent Guenter, and Aude Billard. 2007. On Learning, Representing, and Generalizing a Task in a Humanoid Robot. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)* 37, 2 (2007), 286–298. doi:10.1109/TSMCB.2006.886952

[20] Yuanzhi Cao, Zhuangying Xu, Fan Li, Wentao Zhong, Ke Huo, and Karthik Ramani. 2019. V.Ra: An In-Situ Visual Authoring System for Robot-IoT Task Planning with Augmented Reality. In *Proceedings of the 2019 on Designing Interactive Systems Conference (DIS '19)*. 1059–1070. doi:10.1145/3322276.3322278

[21] Wesley P. Chan, Geoffrey Hanks, Maram Sakr, Haomiao Zhang, Tiger Zuo, H. F. Machiel Van der Loos, and Elizabeth Croft. 2022. Design and Evaluation of an Augmented Reality Head-mounted Display Interface for Human Robot Teams Collaborating in Physically Shared Manufacturing Tasks. *ACM Transactions on Human-Robot Interaction (THRI)* 11, 3 (2022), 1–19. doi:10.1145/3524082

[22] Jiahao Chen, D. Antony Chacon, Muhammad Bilal, Qiushi Zhou, and Wafa Johal. 2024. Mr.LfD: A Mixed Reality Interface for Robot Learning from Demonstration. In *Proceedings of the 36th Australasian Conference on Human-Computer Interaction (OzCHI '24)*. 275–285. doi:10.1145/3726986.3727004

[23] Jason Chen and Alex Zelinsky. 2003. Programing by Demonstration: Coping with Suboptimal Teaching Actions. *The International Journal of Robotics Research* 22, 5 (2003), 299–319. doi:10.1177/0278364903022005002

[24] Mohamed Chetouani. 2021. Interactive Robot Learning: An Overview. *ECCAI Advanced Course on Artificial Intelligence* (2021), 140–172. doi:10.1007/978-3-031-24349-3_9

[25] Nazli Cila, Cristina Zaga, and Maria Luce Lupetti. 2021. Learning from robotic artefacts: A quest for strong concepts in Human-Robot Interaction. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference (DIS '21)*. 1356–1365. doi:10.1145/3461778.3462095

[26] Shanna R. Daly, Colleen M. Seifert, Seda Yilmaz, and Richard Gonzalez. 2016. Comparing Ideation Techniques for Beginning Designers. *Journal of Mechanical Design* 138, 10 (2016). doi:10.1115/1.4034087

[27] Jéssica de Assis Dornelles, Néstor F. Ayala, and Alejandro G. Frank. 2023. Collaborative or substitutive robots? Effects on workers' skills in manufacturing activities. *International Journal of Production Research* 61, 22 (2023), 7922–7955. doi:10.1080/00207543.2023.2240912

[28] K. L. Doty, C. Melchiorri, E. M. Schwartz, and C. Bonivento. 1995. Robot manipulability. *IEEE Transactions on Robotics and Automation* 11, 3 (Jun 1995), 462–468. doi:10.1109/70.388791

[29] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* 41, 4 (2009), 1149–1160. doi:10.3758/BRM.41.4.1149

[30] Sarah E. Fox, Vera Khovanskaya, Clara Crivellaro, Niloufar Salehi, Lynn Dombrowski, Chinmay Kulkarni, Lilly Irani, and Jodi Forlizzi. 2020. Worker-Centered Design: Expanding HCI Methods for Supporting Labor. In *Extended Abstracts of*

the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20). 1–8. doi:10.1145/3334480.3375157

[31] Carl Benedikt Frey and Michael A. Osborne. 2017. The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change* 114 (2017), 254–280. doi:10.1016/j.techfore.2016.08.019

[32] Antoni Grau, Marina Indri, Lucia Lo Bello, and Thilo Sauter. 2020. Robots in Industry: The Past, Present, and Future of a Growing Collaboration With Humans. *IEEE Industrial Electronics Magazine* 15, 1 (2020), 50–61. doi:10.1109/MIE.2020.3008136

[33] Daniel H. Grollman and Aude G. Billard. 2012. Robot Learning from Failed Demonstrations. *International Journal of Social Robotics* 4, 4 (2012), 331–342. doi:10.1007/s12369-012-0161-z

[34] Soheil Habibian, Antonio Alvarez Valdivia, Laura H. Blumenschein, and Dylan P. Losey. 2025. A survey of communicating robot learning during human-robot interaction. *The International Journal of Robotics Research* 44, 4 (2025), 665–698. doi:10.1177/02783649241281369

[35] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology* 52 (1988), 139–183. doi:10.1016/S0166-4115(08)62386-9

[36] Brendan Hertel and S. Reza Ahmadzadeh. 2021. Learning from Successful and Failed Demonstrations via Optimization. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 7807–7812. doi:10.1109/IROS51168.2021.9636679

[37] Kui Hu, Jiwen Zhang, and Dan Wu. 2024. Learning from demonstration for 7-DOF anthropomorphic manipulators without offset via analytical inverse kinematics. *Neurocomputing* 598 (2024), 128036. doi:10.1016/j.neucom.2024.128036

[38] Philipp Jennes and Alberto Di Minin. 2023. Cobots in SMEs: Implementation Processes, Challenges, and Success Factors. In *2023 IEEE International Conference on Technology and Entrepreneurship (ICTE)*. 80–85. doi:10.1109/ICTE58739.2023.10488658

[39] Xinkai Jiang, Paul Mattes, Xiaogang Jia, Nicolas Schreiber, Gerhard Neumann, and Rudolf Lioutikov. 2024. A Comprehensive User Study on Augmented Reality-Based Data Collection Interfaces for Robot Learning. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24)*. 333–342. doi:10.1145/3610977.3634995

[40] Nathan Koenig, Leila Takayama, and Maja Matarić. 2010. Communication and knowledge sharing in human–robot interaction and learning from demonstration. *Neural Networks* 23, 8-9 (2010), 1104–1112. doi:10.1016/j.neunet.2010.06.005

[41] Yue Shin Koh, Gilles Bailly, and Ignacio Avellino. 2025. How Do People Position Assistive Information During Physical Tasks? Insights for Designing Cross-Domain AR Systems. In *Proceedings of the 36th Conference on l'Interaction Humain-Machine (IHM '25)*. 1–14. doi:10.1145/3765712.3765733

[42] David A. Kolb. 1984. *Experiential Learning: Experience as the Source of Learning and Development*.

[43] Przemyslaw A. Lasota, Terrence Fong, and Julie A. Shah. 2017. A Survey of Methods for Safe Human-Robot Interaction. *Foundations and Trends in Robotics* 5, 4 (2017), 261–349. doi:10.1561/2300000052

[44] Jean Lave and Etienne Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press. doi:10.1017/CBO9780511815355

[45] Matthew V. Law, JiHyun Jeong, Amritansh Kwatra, Malte F. Jung, and Guy Hoffman. 2019. Negotiating the Creative Space in Human-Robot Collaborative Design. In *Proceedings of the 2019 on Designing Interactive Systems Conference (DIS '19)*. 645–657. doi:10.1145/3322276.3322343

[46] Christine P. Lee, Pragathi Praveena, and Bilge Mutlu. 2024. REX: Designing User-centered Repair and Explanations to Address Robot Failures. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference (DIS '24)*. 2911–2925. doi:10.1145/3643834.3661559

[47] James R. Lewis. 2018. The System Usability Scale: Past, Present, and Future. *International Journal of Human-Computer Interaction* 34, 7 (2018), 577–590. doi:10.1080/10447318.2018.1455307

[48] Xueyin Li, Xinkai Jiang, Philipp Dahlinger, Gerhard Neumann, and Rudolf Lioutikov. 2025. Beyond Visuals: Investigating Force Feedback in Extended Reality for Robot Data Collection. (2025). doi:10.48550/arXiv.2503.20714

[49] Zhipeng Li, Christoph Gebhardt, Yves Inglin, Nicolas Steck, Paul Streli, and Christian Holz. 2024. SituationAdapt: Contextual UI Optimization in Mixed Reality with Situation Awareness via LLM Reasoning. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*. 1–13. doi:10.1145/3654777.3676470

[50] Rasmus Skovhus Lunding, Tiare Feuchtner, and Kaj Grønbæk. 2025. Investigating AR Assistance for Human-Robot Collaboration in Mould Assembly "in the Wild". In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 540–549. doi:10.1109/HRI61500.2025.10973920

[51] Karl F. MacDorman, Sandosh K. Vasudevan, and Chin-Chang Ho. 2009. Does Japan really have robot mania? Comparing attitudes by implicit and explicit measures. *AI & Society* 23, 4 (2009), 485–510. doi:10.1007/s00146-008-0181-2

[52] Bernardo Marques, Gonçalo Junqueira, João Alves, and Eurico Pedrosa. 2024. Mobile Robots Meet Augmented Reality Technologies: Transforming Human-Robot Interaction in Industry 4.0 Scenarios. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24)*. 740–744. doi:10.1145/3610978.3640681

[53] Ana Moya, Leire Bastida, Pablo Aguirrezabal, Matteo Pantano, and Patricia Abril-Jiménez. 2023. Augmented Reality for Supporting Workers in Human–Robot Collaboration. *Multimodal Technologies and Interaction* 7, 4 (2023), 40. doi:10.3390/mti7040040

[54] Tamara Munzner. 2014. *Visualization Analysis and Design*. doi:10.1201/b17511

[55] D. Ni, A. W. W. Yew, S. K. Ong, and A. Y. C. Nee. 2017. Haptic and visual augmented reality interface for programming welding robots. *Advances in Manufacturing* 5, 3 (2017), 191–198. doi:10.1007/s40436-017-0184-7

[56] Soh Khim Ong, A. W. W. Yew, Naresh Kumar Thanigaivel, and Andrew Y. C. Nee. 2020. Augmented reality-assisted robot programming system for industrial applications. *Robotics and Computer-Integrated Manufacturing* 61 (2020), 101820. doi:10.1016/j.rcim.2019.101820

[57] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J. Andrew Bagnell, Pieter Abbeel, and Jan Peters. 2018. An Algorithmic Perspective on Imitation Learning. *Foundations and Trends in Robotics* 7, 1-2 (2018), 1–179. doi:10.1561/2300000053

[58] Sun Young Park, Pei-Yi Kuo, Andrea Barbarin, Elizabeth Kaziunas, Astrid Chow, Karandeep Singh, Lauren Wilcox, and Walter S. Lasecki. 2019. Identifying Challenges and Opportunities in Human-AI Collaboration in Healthcare. In *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing (CSCW '19)*. 506–510. doi:10.1145/3311957.3359433

[59] Sharon K Parker, Timothy Ballard, Mark Billinghurst, Catherine Collins, Maureen Dollard, Mark A Griffin, Wafa Johal, Karina Jorritsma, Marek Kowalkiewicz, Eva Kyndt, Dean Lusher, Char lee McLennan, Tim Miller, Andrew Neal, Jeannie Marie Paterson, Frank Vetere, and Toby Walsh. 2025. Quality work in the future: New directions via a co-evolving sociotechnical systems perspective. *Australian Journal of Management* (2025). doi:10.1177/03128962251331813

[60] Vaidehi Patil, Jyotindra Narayan, Kamalpreet Sandhu, and Santosha K. Dwivedy. 2022. Integration of Virtual Reality and Augmented Reality in Physical Rehabilitation: A State-of-the-Art Review. *Revolutions in Product Design for Healthcare: Advances in Product Design and Design Methods for Healthcare* (2022), 177–205. doi:10.1007/978-981-16-9455-4_10

[61] Ornnalin Phaijit, Claude Sammut, and Wafa Johal. 2023. User Interface Interventions for Improving Robot Learning from Demonstration. In *Proceedings of the 11th International Conference on Human-Agent Interaction (HAI '23)*. 152–161. doi:10.1145/3623809.3623848

[62] Harish Ravichandar, Athanasios S. Polydoros, Sonia Chernova, and Aude Billard. 2020. Recent Advances in Robot Learning from Demonstration. *Annual Review of Control, Robotics, and Autonomous Systems* 3 (2020), 297–330. doi:10.1146/annurev-control-100819-063206

[63] Eric Rosen, David Whitney, Michael Fishman, Daniel Ullman, and Stefanie Tellex. 2020. Mixed Reality as a Bidirectional Communication Interface for Human-Robot Interaction. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 11431–11438. doi:10.1109/IROS45743.2020.9340822

[64] Farshad Safavi, Parthan Olikkal, Dingyi Pei, Sadia Kamal, Helen Meyerson, Varsha Penumalee, and Ramana Vinjamuri. 2024. Emerging Frontiers in Human–Robot Interaction. *Journal of Intelligent & Robotic Systems* 110, Article 45 (2024). doi:10.1007/s10846-024-02074-7

[65] Maram Sakr, Zexi Jesse Li, H. F. Machiel Van der Loos, Dana Kulić, and Elizabeth A. Croft. 2022. Quantifying Demonstration Quality for Robot Learning and Generalization. *IEEE Robotics and Automation Letters* 7, 4 (2022), 9659–9666. doi:10.1109/LRA.2022.3191950

[66] Maram Sakr, Zhikai Zhang, Benjamin Li, Haomiao Zhang, H. F. Machiel Van der Loos, Dana Kulić, and Elizabeth Croft. 2025. How Can Everyday Users Efficiently Teach Robots by Demonstration? *ACM Transactions on Human-Robot Interaction* 14, 4, Article 74 (2025), 22 pages. doi:10.1145/3737892

[67] Sarah Schömbs, Jorge Goncalves, and Wafa Johal. 2025. "I can feel the risks by looking at the robot face": Communicating Risk through a Physical Agent. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25)*. 236–252. doi:10.1145/3715336.3735759

[68] Mariah L. Schrum, Erin Hedlund-Botti, Nina Moorman, and Matthew C. Gombolay. 2022. MIND MELD: Personalized Meta-Learning for Robot-Centric Imitation Learning. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Vol. 2022-March. 157–165. doi:10.1109/HRI53351.2022.9889616

[69] Aran Sena and Matthew Howard. 2020. Quantifying teaching behavior in robot learning from demonstration. *The International Journal of Robotics Research* 39, 1 (2020), 54–72. doi:10.1177/0278364919884623

[70] Pavel Senin. 2008. Dynamic Time Warping Algorithm Review. *Technical Report CSDL-08-04* (2008), 23 pages.

[71] Adwait Sharma, Alexander Ivanov, Frances Lai, Tovi Grossman, and Stephanie Santosa. 2024. GraspUI: Seamlessly Integrating Object-Centric Gestures within the Seven Phases of Grasping. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference (DIS '24)*. 1275–1289. doi:10.1145/3643834.3661551

[72] Bruno Siciliano and Oussama Khatib. 2016. Robotics and the Handbook. In *Springer Handbook of Robotics*. Springer. doi:10.1007/978-3-319-32552-1_1

[73] Laura Stegner, Yuna Hwang, David Porfirio, and Bilge Mutlu. 2024. Understanding On-the-Fly End-User Robot Programming. In *Proceedings of the 2024*

ACM Designing Interactive Systems Conference (DIS '24). 2468–2480. doi:10.1145/3643834.3660721
[74] John Sweller. 1988. Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science* 12, 2 (1988), 257–285. doi:10.1207/s15516709cog1202_4
[75] Kiichiro Tatsuzawa, Jiayi Wu, Jo Eann Chong, Wenyuan Wu, Qiting Ma, David Kelly, Jessica Stander, Alireza Mohammadi, Arzoo Atiq, and D. Antony Chacon. 2025. vARtebrae: Medical Simulation for Spinal Mobilisation Employing Mechanical Metamaterials and Extended Reality. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. 1–7. doi:10.1145/3706599.3719996
[76] Annalisa T. Taylor, Thomas A. Berrueta, and Todd D. Murphey. 2021. Active learning in robotics: A review of control principles. *Mechatronics* 77 (2021), 102576. doi:10.1016/j.mechatronics.2021.102576
[77] Gareth Terry and Nikki Hayfield. 2021. *Essentials of Thematic Analysis*. American Psychological Association. doi:10.1037/0000238-000
[78] Kristina Tornbjerg, Anne Marie Kanstrup, Mikael B. Skov, and Matthias Rehm. 2021. Investigating human-robot cooperation in a hospital environment: Scrutinising visions and actual realisation of mobile robots in service work. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference (DIS '21)*. 381–391. doi:10.1145/3461778.3462101
[79] Ana-Lucia Pais Ureche and Aude Billard. 2015. Metrics for Assessing Human Skill When Demonstrating a Bimanual Task to a Robot. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*. 37–38. doi:10.1145/2701973.2702017
[80] Daryoush Vaziri, David Golchinfar, Gunnar Stevens, and Dirk Schreiber. 2020. Exploring Future Work - Co-Designing a Human-robot Collaboration Environment for Service Domains. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference (DIS '20)*. 153–164. doi:10.1145/3357236.3395483
[81] Steven Vi, Tiago Silva da Silva, and Frank Maurer. 2019. User Experience Guidelines for Designing HMD Extended Reality Applications. In *IFIP Conference on Human-Computer Interaction*. 319–341. doi:10.1007/978-3-030-29390-1_18
[82] Jan Ole von Hartz, Tim Welschehold, Abhinav Valada, and Joschka Boedecker. 2024. The Art of Imitation: Learning Long-Horizon Manipulation Tasks From Few Demonstrations. *IEEE Robotics and Automation Letters* 9, 12 (2024), 11369–11376. doi:10.1109/LRA.2024.3487506
[83] L. S. Vygotsky. 1978. *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press. doi:10.2307/j.ctvjf9vz4
[84] Michael Walker, Thao Phung, Tathagata Chakraborti, Tom Williams, and Daniel Szafir. 2023. Virtual, Augmented, and Mixed Reality for Human-robot Interaction: A Survey and Virtual Design Element Taxonomy. *ACM Transactions on Human-Robot Interaction* 12, 4 (2023), 1–39. doi:10.1145/3597623
[85] Keru Wang, Zhu Wang, Ken Nakagaki, and Ken Perlin. 2024. "Push-That-There": Tabletop Multi-robot Object Manipulation via Multimodal 'Object-level Instruction'. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference (DIS '24)*. 2497–2513. doi:10.1145/3643834.3661542
[86] Xian Wang, Luyao Shen, and Lik-Hang Lee. 2025. A Systematic Review of XR-Enabled Remote Human-Robot Interaction Systems. *ACM Computing Surveys* 57, 11 (2025), 1–37. doi:10.1145/3730574
[87] Colin Ware. 2019. *Information Visualization: Perception for Design* (4th ed.).
[88] David Wood, Jerome S. Bruner, and Gail Ross. 1976. The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry* (1976). doi:10.1111/j.1469-7610.1976.tb00381.x
[89] Maciej K. Wozniak, Max Pascher, Bryce Ikeda, Matthew B. Luebbers, and Ayesha Jena. 2024. Virtual, Augmented, and Mixed Reality for Human-Robot Interaction (VAM-HRI). In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 1361–1363. doi:10.1145/3610978.3638158
[90] Tsuneo Yoshikawa. 1985. Manipulability of Robotic Mechanisms. *The International Journal of Robotics Research* 4, 2 (1985), 3–9. doi:10.1177/027836498500400201
[91] Qiushi Zhou, Antony Chacon, Jiahe Pan, and Wafa Johal. 2025. Assisting MoCap-Based Teleoperation of Robot Arm using Augmented Reality Visualisations. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 1765–1769. doi:10.1109/HRI61500.2025.10974111
[92] Qiushi Zhou, D. Antony Chacon, Jiahe Pan, and Wafa Johal. 2026. Ghost Arm: Aligning Human and Robot Kinematics through AR Overlays in MoCap-Based Teleoperation of Robot Arm. In *Proceedings of the Augmented Humans International Conference 2026 (AHs '26)*. 12 pages. doi:10.1145/3795011.3795014

Appendix

A.1 Performance Metrics

To quantify how each experimental condition affects both human performance and the resulting robot performance, we define a set of objective metrics based on the robot's kinematic state. Specifically, consider a robot with n degrees of freedom, whose joint variables are denoted by $\mathbf{q}(t) = [q_1(t), \ldots, q_n(t)]$, where $q_i \in \mathbb{R}$. A demonstration D is defined as a discrete joint-space trajectory:

$$D = \langle \mathbf{q}(t_0), \mathbf{q}(t_1), \ldots, \mathbf{q}(t_T) \rangle, \tag{1}$$

where $\mathcal{T}_m = \{t_0, \ldots, t_T\}$ denotes the execution interval of the mth demonstration, and $|\mathcal{T}_m|$ is the total number of recorded time steps. Let $\mathcal{D} = \langle D_1, D_2, \ldots, D_{\mathcal{M}} \rangle$ denote the set of all given demonstrations, where the total number of demonstrations is represented as $\mathcal{M} = |\mathcal{D}|$. All metrics below are computed only over successful demonstrations $\mathcal{D}_s \subset \mathcal{D}$.

Task Completion Rate — Defined as the proportion of demonstrations that successfully achieved the task objective, computed as the ratio of successful demonstrations $|\mathcal{D}_s|$ to the total number of demonstrations $\mathcal{M}$.

Active Demonstration Time — To quantify the effective duration of human demonstrations, we computed the *active demonstration time* by excluding periods in which the robot remained idle. Joint velocities were obtained by finite differencing the recorded joint positions at 30 Hz. At each timestep, the robot was classified as idle if the maximum absolute joint velocity across all joints was below a motion threshold (0.017 rad/s), reflecting encoder noise bounds and the minimum intentional guiding force under impedance control. Idle segments shorter than 0.2 s were discarded to avoid spurious detections. The total idle duration was subtracted from the total demonstration duration to yield the active demonstration time.

Distance to Joint Limits — To assess how closely the robot operated to its mechanical limits, we computed a normalized distance-to-limit measure. For each joint i with limits $q_{i,\min}$ and $q_{i,\max}$, the normalized clearance at time t is

$$d_i(t) = \frac{\min\big(q_i(t) - q_{i,\min},\ q_{i,\max} - q_i(t)\big)}{q_{i,\max} - q_{i,\min}}. \tag{2}$$

We then averaged across all n joints to obtain a demonstration-level profile:

$$d_{\text{joint}}(t) = \frac{1}{n} \sum_{i=1}^{n} d_i(t), \tag{3}$$

and computed the mean clearance over the full trajectory:

$$d^m_{\text{joint}} = \frac{1}{|\mathcal{T}_m|} \sum_{t \in \mathcal{T}_m} d_{\text{joint}}(t), \tag{4}$$

where higher values of d^m_{joint} indicate that the robot operated farther from its joint limits.

Distance to Self-Collision — We quantified the safety margin of robot motion with respect to internal collisions using the MoveIt 2 collision environment. For each configuration $\mathbf{q}(t) \in D^m$, the robot state was inserted into the planning scene and signed distances were computed between all non-ignored link pairs. For links i and j, let $d_{ij}(t)$ denote their signed Euclidean separation, where negative values indicate penetration (self-collision) configurations and positive values indicate collision-free separation. The instantaneous minimum separation is

$$d_{\min}(t) = \min_{(i,j)} d_{ij}(t). \tag{5}$$

Table A1: NASA-TLX Questionnaire Results (21-point Scale) — $M \pm SD$ with 95% Confidence Intervals.

		Offline Feedback	Real-Time Feedback	Combined Feedback
Insertion	Mental Demand	6.92 ± 3.63 [4.61, 9.22]	6.17 ± 4.76 [3.14, 9.19]	6.67 ± 1.67 [5.61, 7.73]
	Physical Demand	8.50 ± 3.80 [6.08, 10.92]	8.00 ± 4.51 [5.13, 10.87]	6.67 ± 1.44 [5.75, 7.58]
	Temporal Demand	6.75 ± 3.41 [4.58, 8.92]	5.83 ± 3.35 [3.70, 7.96]	6.67 ± 2.06 [5.36, 7.98]
	Performance	12.33 ± 5.53 [8.82, 15.85]	16.00 ± 2.83 [14.20, 17.80]	16.75 ± 2.86 [14.93, 18.57]
	Effort	8.83 ± 3.13 [6.85, 10.82]	6.67 ± 3.52 [4.43, 8.91]	6.50 ± 1.93 [5.27, 7.73]
	Frustration	6.75 ± 4.09 [4.15, 9.35]	4.42 ± 3.63 [2.11, 6.72]	5.50 ± 2.11 [4.16, 6.84]
Pouring	Mental Demand	12.25 ± 4.59 [9.33, 15.17]	6.08 ± 5.02 [2.90, 9.27]	6.42 ± 4.06 [3.84, 8.99]
	Physical Demand	11.00 ± 4.37 [8.22, 13.78]	5.92 ± 5.20 [2.62, 9.22]	7.92 ± 3.82 [5.49, 10.35]
	Temporal Demand	7.75 ± 4.05 [5.18, 10.32]	4.83 ± 3.35 [2.70, 6.96]	6.67 ± 2.93 [4.80, 8.53]
	Performance	8.75 ± 5.15 [5.48, 12.02]	15.50 ± 3.85 [13.05, 17.95]	15.33 ± 3.03 [13.41, 17.26]
	Effort	14.00 ± 4.75 [10.98, 17.02]	6.17 ± 4.95 [3.02, 9.31]	8.83 ± 3.86 [6.38, 11.28]
	Frustration	11.67 ± 4.54 [8.78, 14.55]	4.00 ± 4.20 [1.33, 6.67]	4.75 ± 4.43 [1.93, 7.57]

A conservative safety margin for demonstration m is obtained as the minimum distance over its execution interval:

$$d^m_{\text{self}} = \min_{t \in \mathcal{T}_m} d_{\min}(t), \tag{6}$$

where larger values of d^m_{self} correspond to safer kinematic configurations during the demonstration.

Quality of Demonstration (Manipulability) — To evaluate the dexterity of demonstrated configurations, we used the manipulability measure [90], which reflects proximity to kinematic singularities and has been linked to demonstration quality in learning-from-demonstration settings [62]. For an x-dimensional task with Jacobian $J(\mathbf{q}(t)) \in \mathbb{R}^{x \times n}$, the manipulability at time t for demonstration m is

$$\Omega(D^m, t) = \sqrt{\det\big(J(\mathbf{q}(t))J^\top(\mathbf{q}(t))\big)}. \tag{7}$$

The demonstration-level manipulability score is obtained by averaging over the trajectory:

$$\alpha^m = \frac{1}{|\mathcal{T}_m|} \sum_{t \in \mathcal{T}_m} \Omega(D^m, t). \tag{8}$$

Higher α^m values indicate more dexterous and higher-quality demonstrations, whereas lower values reflect reduced motion flexibility.

A.2 Policy Learning

For the learning algorithm, we used a standard Gaussian Mixture Model (GMM) to train LfD models from the collected data [17, 19]. This approach is widely used in LfD research due to its ability to capture the underlying distribution of demonstrated trajectories from a limited number of demonstrations with minimal learning time [57, 69]. The robot task learning process begins after the user provides a set of $\mathcal{M}$ demonstrations, where each demonstration m consists of T_m state data points. Each state s is represented as $\xi_s = (t_s, x_s^T, \omega_s^T)$, where t_s, x_s, and ω_s correspond to the time, end-effector position, and orientation in quaternion form, respectively.

After temporally aligning the demonstrations, a K-component GMM is fitted to all demonstration data, with parameters (π_k, μ_k, Σ_k) representing the prior probabilities, means, and covariance matrices for each component k. The number of components K is selected using the Bayesian Information Criterion (BIC) to balance model complexity and data fit. For trajectory generation, Gaussian Mixture Regression (GMR) is applied to predict the expected end-effector position and orientation at each time step. This approach allows the robot to reproduce smooth and generalized trajectories from the given demonstrations [17].

A.3 Interview Questions

Q1: What changes did you make after observing the robot's execution?
Q2: What adjustments did you make after failed demonstrations (if any)?
Q3: Did you understand the key qualities (or what features were you paying attention to) that make a demonstration effective?
Q4: What would you improve if you were to repeat the task?
Q5: How did you perceive the impact of your demonstrations on the robot's performance?
Q6: How did the feedback system help you perform the second task effectively?

A.4 Focus Group Drawings

Figures A1–A3 summarize the group-level "best" drawings from the focus group study, illustrating participant-generated AR visualizations for joint limits, self-collision, and manipulability.

Figure A1: Group-level best drawings for joint limits. Participants proposed joint-limit communication in AR using a bottom-up hierarchy that reflects increasingly sophisticated feedback layers. At the foundation, they emphasized intuitive awareness cues, especially color-based indicators that communicate safety at a glance. Building on this, they highlighted the need to quantify proximity to mechanical limits for more precise control. Participants also stressed placing these cues spatially on or near the joint. As the hierarchy progresses, alert mechanisms should escalate as users approach or exceed critical thresholds. Finally, at the highest level, participants emphasized the importance of customization.

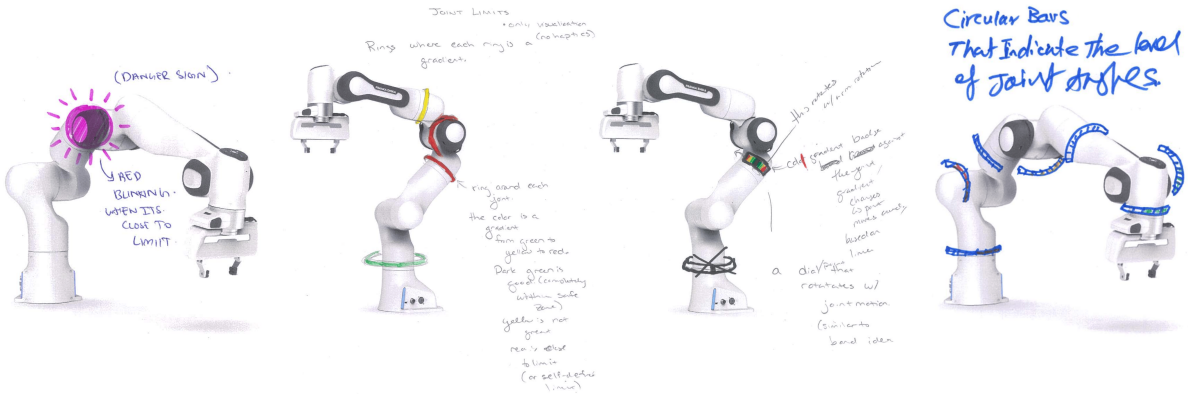


Figure A2: Group-level best drawings for self-collision. Participants described a hierarchy of needs for managing robot self-collisions in AR, starting with basic collision avoidance and progressing toward deeper spatial understanding. At the foundation, users need clear indicators of collision proximity to detect potential contact. Building on this, participants emphasized *actionable guidance* to help users recover safely. They also noted that the form of feedback, from simple 2D cues to richer 3D meshes, affects how well users interpret spatial relationships and anticipate collisions. Finally, they highlighted the need to manage attentional load by distinguishing normal states from genuine danger to avoid overwhelming users.

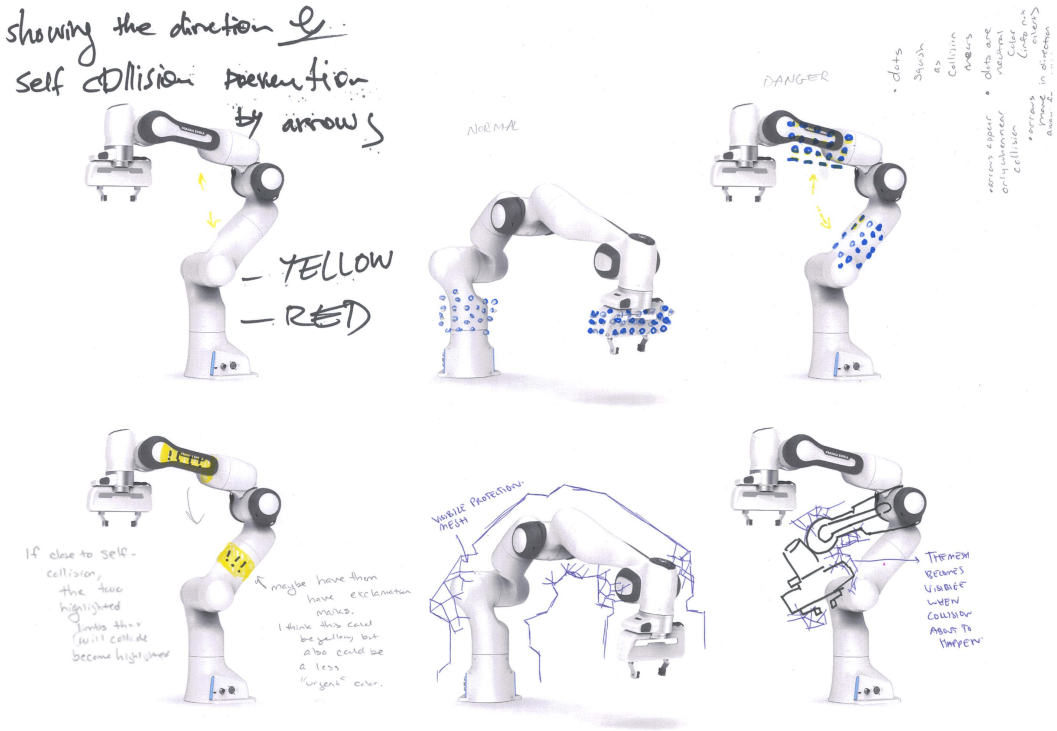


Figure A3: Group-level best drawings for manipulability. Participants described a layered set of needs for supporting manipulability in AR during robot learning from demonstration. First, they emphasized presenting information in the right visual form and location so users can distinguish between local joint issues and global posture effects. They then highlighted the need for clear alerts when the robot approaches poor manipulability. Building on this, participants stressed *actionable guidance*, such as directional cues or preventive mechanisms, to help users adjust demonstrations and avoid singular configurations.