RFM-HRI : A Multimodal Dataset of Medical Robot Failure, User Reaction and Recovery Preferences for Item Retrieval Tasks

While robots deployed in real-world environments inevitably experience interaction failures, understanding how users respond through verbal and non-verbal behaviors remains under-explored in human–robot interaction (HRI). This gap is particularly significant in healthcare-inspired settings, where interaction failures can directly affect task performance and user trust. We present the Robot Failures in Medical HRI (RFM-HRI) Dataset, a multimodal dataset capturing dyadic interactions between humans and robots embodied in crash carts, where communication failures are systematically induced during item retrieval tasks.

Through Wizard-of-Oz studies with 41 participants across laboratory and hospital settings, we recorded responses to four failure types (speech, timing, comprehension, and search) derived from three years of crash-cart robot interaction data. The dataset contains 214 interaction samples including facial action units, head pose, speech transcripts, and post-interaction self-reports. Our analysis shows that failures significantly degrade affective valence and reduce perceived control compared to successful interactions.

Failures are strongly associated with confusion, annoyance, and frustration, while successful interactions are characterized by surprise, relief, and confidence in task completion. Emotional responses also evolve across repeated failures, with confusion decreasing and frustration increasing over time.

This work contributes (1) a publicly available multimodal dataset (RFM-HRI), (2) analysis of user responses to different failure types and preferred recovery strategies, and (3) a crash-cart retrieval scenario enabling systematic comparison of recovery strategies with implications for safety-critical failure recovery. Our findings provide a foundation for failure detection and recovery methods in embodied HRI.

Additional Key Words and Phrases: Human-Robot Interaction; Robot Failure and Recovery; Multimodal Datasets; Affective Response; Healthcare Robotics

1 Introduction

The integration of autonomous systems into real-world deployments has highlighted the inevitability of interaction failure. When robots fail, humans inherently respond both verbally (i.e., speech [15]) and non-verbally (i.e., gestures, facial expressions [11], posture etc). However, the manner in which people respond to the robot may depend on the robot's use of verbal or non-verbal communicative cues during failures, how the robot chooses to recover from such failures, and how appropriate this response is. [1, 27]. Despite this, limited work in HRI has explored methods by which robots can effectively recover from failures using multimodal cues in real-time [15, 31].

To date, HRI research has primarily examined robot failures through asynchronous video evaluations [28] and in social environments [25], with limited attention to workplace settings where stressful conditions place different demands on users. As a result, the HRI community could benefit from failure datasets that capture technical robot failures and user reactions related to safety critical scenarios.

Healthcare represents a particularly demanding safety-critical domain, characterized by extreme time pressure, rapidly changing conditions, and high cognitive load [18]. During emergency procedures, delays or errors in accessing medications and supplies can disrupt workflow and divert attention away from patient care and team coordination, with outsized consequences for team performance and timely intervention [7].

In emergency medicine, access to supplies is often mediated through shared physical artifacts such as crash carts, which centralize critical medications and equipment [18]. Locating items on a crash cart is a distributed, collaborative activity that requires interpreting verbal requests, maintaining awareness of cart organization, and coordinating actions

Author's Contact Information:

Fig. 1. Data collection environment at university campus laboratory (left) and at the hospital (right) Participants faces are blacked out in compliance with our approved IRB protocol.

under stress [35]. Furthermore, the frequency of crash cart use varies by hospital unit; as a result, some healthcare workers interact with the cart far less frequently and may struggle to recall where specific supplies are located. Our long-term vision is to design a crash cart robot that supports efficient access to critical supplies, reducing delays in patient care.

Within time-critical and coordination-intensive tasks, robot errors can be categorized from two perspectives: the technical system perspective, which focuses on deviations from the robot's designed functionality, and the user perspective, which centers on deviations from human expectations and impacts on interaction. From the user perspective, failures tackle the problem of detecting robot errors using affect, gaze, or posture [14, 17]. Complementary research models error detection from the system-perspective through norm violation or anomaly detection frameworks [26].

While a growing body of work explores robot recovery behaviors, these efforts focus almost entirely on verbal repair and conversational agents [2]. Datasets such as the ERR@HRI 2.0 Challenge [9, 30], REFLEX [19], REACT [8], Response-to-Errors [32], and EMPATHIC [10] have propelled research interest in this domain, by providing multimodal recordings of conversations, social signals, and user reactions.

Despite this progress in socially grounded and conversational settings, several gaps remain. First, because many robot failure modes studied in HRI are designed for social settings, their applicability to safety-critical workplace environments, such as healthcare, remains unclear. Second, existing HRI datasets do not provide paired annotations of failure events and user-preferred recovery behaviors in physically grounded healthcare interactions. Third, even when recovery strategies are studied, their preference structure across distinct failure types has not been systematically examined in safety-critical contexts.

These challenges motivate the study of assistive robotic systems for rapid, reliable item retrieval in healthcare-oriented tasks. In this paper, we address these gaps by conducting a structured failure-induction study with a crash cart robot that guides users through item retrieval tasks. As part of this study, we first derived four failure types (speech, timing, comprehension, and search) from three years of empirical observation of interactions with crash cart robots in emergency medicine contexts. We then collected a multimodal dataset including 214 samples of HRI in laboratory

and hospital environments, which capture synchronized facial expressions, head pose, speech, and post-interaction reflections on recovery strategy preferences (Figure 1).

The item retrieval task captures core elements of general item retrieval problems in embodied robotics, including goal specification, spatial search, manipulation, communication, and recovery from failure. These challenges are shared across a range of real-world deployments, such as manufacturing and factory settings, where robots must reliably retrieve, transport, and hand off objects under time pressure and partial observability. As such, RFM-HRI can support research on item retrieval, failure detection, and recovery strategy learning beyond healthcare domains.

Our **contributions are threefold**. **First**, we introduce RFM-HRI, a novel multimodal dataset of dyadic interactions with a crash cart robot, which will be made publicly available to increase research interest in evaluating robot failures and recovery strategies. The dataset has 214 samples from 41 participants across laboratory and hospital settings, each including synchronized facial action units, head pose, speech transcriptions, and post-interaction self reports of emotional reactions and robot recovery preferences. **Second**, we design and evaluate a set of robot communication failure modes, including speech, timing, search, and comprehension failures, and analyze how users respond to each failure type and its associated robot recovery strategy. Failure trials are associated with more negative emotional responses and reduced sense of control relative to successful trials, with confusion, annoyance, and frustration dominating reported emotions. Emotional patterns vary across failure types and over repeated interactions. We also document a strong preference for verbal recovery strategies, with multimodal and nonverbal strategies selected far less frequently. **Third**, we contribute a novel HRI failure scenario centered on guided search and item retrieval tasks in a crash cart robot, providing a realistic interaction for studying failure and recovery behavior in healthcare and other structured item retrieval settings. Through a systematic comparison of verbal and non-verbal robot recovery strategies, we offer design implications for robots that engage in failure to recovery interaction loops in safety-critical HRI settings.

2 Related Work

2.1 Conceptual Foundations

An error is the internal manifestation or state change resulting from the fault (e.g., the robot's navigation system calculates an incorrect internal position). A failure is the ultimate external event where the system deviates from its expected service, becoming observable to the human user (e.g., the robot bumps into a wall or violates a social norm by interrupting) [20].

2.1.1 Definition of Robot Failures. Based on the observations of the robot errors from robot field studies over the course of 3 years [33–35], we identified four communication errors that occur during HRIs in the emergency room. Before describing these specific failures, we clarify two complementary perspectives for understanding robot errors in HRI.

2.1.2 Error Perspectives. We distinguish between two perspectives for analyzing robot failures, system and user perspective errors.

System Perspective Errors are failures or undesired behaviors that originate from the robot's internal components or algorithms, such as a sensor malfunction or delayed speech output. These are considered robot-side faults when viewed from the developer's perspective, and represent technical breakdowns in the robot's designed functionality.

User Perspective Errors are failures that are experienced by the human user due to a mismatch, miscommunication, or human misunderstanding. These are failures that are perceived in the context of HRI, even if the robot hardware or

software is technically operating correctly. User perspective errors capture how failures manifest in the lived experience of interaction, regardless of their technical root cause.

Differentiating between these perspectives explicitly allows us to label and analyze failure events in terms of both internal root causes (system perspective) and human experience (user perspective). This dual framing is essential for developing robots that can both detect technical malfunctions and recognize when users perceive something has gone wrong. This terminology provides a shared vocabulary for connecting technical breakdowns to how failures are perceived and managed during interaction, which motivates prior HRI models of failure handling.

2.2 Robot Failures

In HRI, robot failures are increasingly treated as interaction events that shape what users notice, how they interpret robot behavior, and what corrective actions they attempt [17]. Honig and Oron-Gilad synthesized prior work on failure handling and introduced the Robot Failure Human Information Processing (RF-HIP) model, which frames failure handling as staged human information processing (e.g., failure communication, perception and comprehension, and problem solving), with each stage influenced by contextual factors and mitigation strategies [17]. They further argued that failure handling has downstream consequences for trust, performance, and willingness to reuse robots, motivating datasets that capture not only failures themselves but also how users experience and respond to them [17].

These concerns are further sharpened in healthcare and emergency settings, where time pressure, safety criticality, and team coordination demands can amplify the consequences of failures [35]. Taylor et al. presented a rapidly built medical crash cart robot deployed in emergency rooms, and introduced a taxonomy of failure modes observed during real-world clinical use, including breakdowns in communication, timing, and coordination. This work highlights how robot failures in healthcare are embedded within broader team workflows and clinical constraints, and underscores the importance of analyzing failure and recovery in embodied medical robots [35].

To date, healthcare robot deployments have largely emphasized qualitative analysis and system-level insights, and fewer efforts have focused on releasing publicly available datasets that combine labeled failure modes with multimodal annotations of user reactions for machine learning research [35].

Our work complements this body of research by pairing structured failure induction with multimodal measurements of user responses in a controlled experimental setting, supporting both human-centered analysis and data-driven modeling of human reactions to robot failures.

2.3 Robot Error Detection Using Social Signals

A complementary line of work studies how *social signals*, including facial expressions, head movements, gaze, prosody, and body posture, can be used to recognize when a robot-assisted task is going wrong [6]. Bremers et al. review this literature and argue that social cues can be early indicators of task failure, occurring before explicit verbal complaint or task abort [6]. They highlight that models leveraging gaze shifts, facial tension, hesitation markers, or changes in speaking rate can distinguish successful from failed interactions, while also noting that many existing studies are small-scale and domain-specific [6].

Recent multimodal work operationalizes this idea in conversational settings [9]. The ERR@HRI 2.0 challenge provides a dataset of dyadic human-robot conversations with LLM-powered robots, annotated for robot errors and users' corrective intent, with accompanying multi modal features[9]. This corpus supports automatic error detection from facial and vocal cues, but its tasks (e.g., trip planning, debates) differ substantially from time-critical medical retrieval and physically grounded human-robot coordination [9].

Most similar to "social-signal response to errors" in physical HRI, the *Response-to-Errors in HRI* dataset curated by Stiber et al. aggregates multiple physical HRI scenarios and annotates temporal boundaries around robot errors and human reactions, with facial action units extracted via OpenFace [32]. In their setup, each entry corresponds to an error episode (pre-error, during-error, reaction, post-error) with annotated timestamps for perceived error start and reaction onset/offset [32]. This setup provides an important precedent for treating reactions as temporally localized signals around induced or unexpected errors in embodied tasks [32].

Our dataset follows the social-signal perspective but in a different setting. We study an embodied crash cart robot that uses sound and lights to guide item retrieval during simulated emergency scenarios, where failures are systematically injected across speech, timing, comprehension, and search [6]. We record videos, including facial expressions (processed with Google MediaPipe-style landmark pipelines), audio (processed with speech-to-text), and post-study survey responses after participants re-watch their interactions, allowing us to relate objective social-signal patterns to both online behavior and retrospective self-reports in a physically grounded task [6].

To situate these signals and failure annotations within the broader landscape of available resources, we next summarize relevant datasets spanning affective computing, conversational error modeling, and embodied HRI.

2.4 HRI Datasets for Failures and Failure Responses

We summarize datasets that capture unintentional events and affect, conversational errors, and embodied HRI failures and reactions, including healthcare-focused multimodal resources (Table 1).

2.4.1 Unintentional Events and Affective Computing. The OOPS! dataset collects in-the-wild videos of unintentional human actions and labels the transition from intentional behavior to accidental "oops" moments [13]. It is a valuable resource for learning to anticipate accidents, but it does not involve robots, healthcare settings, or explicit user reactions to robot behavior [13].

Large facial-expression datasets such as RAF-DB [21], CK+ [23], Oulu-CASIA [37], AFEW 7.0 [12], BP4D-Spontaneous [36], and related corpora provide labels for emotions and/or Facial Action Coding System (FACS) and Action Units (AUs), across posed and in-the-wild (or semi-structured) recordings. For example, Oulu-CASIA is explicitly designed to include both near-infrared (NIR) and visible recordings for basic facial expressions [37], while BP4D-Spontaneous provides high-resolution dynamic facial behavior with frame-level FACS annotations and 3D facial and head tracking metadata, captured under a structured elicitation protocol [36]. These datasets underpin many tools for AU detection and affect modeling, but they are not collected in HRI settings and do not target reactions to task or safety-critical robot failures.

2.4.2 Conversational Error Datasets. Within conversational AI, the ACE dataset annotates conversational agent errors in human-agent interactions to support the analysis of error types and user-perceived breakdowns [3]. ERR@HRI-style corpora extend beyond text to include multimodal behavioral features and temporally localized error/reaction annotations in conversational HRI [9]. However, these datasets focus on conversational tasks and do not capture physical interactions with robots during urgent medical item retrieval tasks [3, 9].

2.4.3 Embodied HRI Failures and Reactions. Within HRI and human-agent interaction, multiple datasets explicitly connect agent/robot mistakes to human reactions. The EMPATHIC [10] framework includes a dataset of human facial reactions while participants *observe* an autonomous agent that executes (often suboptimal) behavior, with structured incentives so that observers care about outcomes. This is valuable for learning reaction-to-quality mappings, but it is

Table 1. Comparison of representative datasets spanning facial expression analysis, conversational errors, and human–robot interaction. In contrast to prior work, **RFM-HRI (Ours)** provides a healthcare-centered HRI dataset featuring a novel crash cart robot, controlled failure injection, and rich multimodal annotations of human reactions, including speech, FACS and AUs, head pose, gaze estimation, and post-hoc survey responses—collected from both expert and non-expert participants.

Dataset	Annotations	Dataset Modalities	Healthcare Domain?	Experts and Non-Experts?	Human Robot Interaction?	Controlled Failure Injection?	Reactions to Failures?
Oops! [13]	Videos Annotated for Action Intention Variability	Video	×	×	×	×	✓
RAF-DB [21]	Compound Expression Images and Labels in the Wild	Image	×	×	×	×	×
CK+ [23]	Facial Action Coding System, Action Units, and Labels for Posed and Unposed Expressions	Video, Image, FACS/AUs	×	×	×	×	×
Oulu-CASIA [37]	Facial Expression Dataset, Across Illumination Conditions	Image (Infrared and Visible Light)	×	×	×	×	×
AFEW 7.0 [12]	Emotion-Labeled Movie Clips	Video, Audio	×	×	×	×	×
BP4D-Spontaneous [36]	2D and 3D Dataset Containing Facial Action Units, Video Features, and Head Poses	Video (2D and 3D), FACS/AUs, Pose	×	×	×	×	×
ACE [3]	Conversational Errors and Analysis in Human-Agent Interaction	Transcripts, Annotations	×	×	×	✓	✓
ERR@HRI 2.0 [9]	Multimodal Dataset of Conversational HRI with System-Failures	FACS/AUs, Pose, Audio Features, Transcripts	×	×	✓	✓	✓
EMPATHIC [10]	Implicit human feedback (facial reactions)	Video, Annotations	×	×	×	✓	✓
REACT [8]	Human reactions & evaluative feedback	FACS/AUs, Pose, Gaze, Annotations	×	×	✓	✓	✓
REFLEX [19]	Human reactions to failures/explanations	FACS/AUs, Pose, Gaze, Transcript, Annotations	×	×	✓	✓	✓
Response-to-Errors [32]	Facial AU social responses to errors	AUs, Annotations	×	×	✓	✓	✓
HRI-SENSE [16]	Multimodal social/emotional responses	Videos (with Depth), FACS/AUs, Pose, Gaze, Transcripts, Survey Data	×	×	✓	✓	✓
OpenRoboCare [22]	Expert caregiving tasks	Video (RGB–D), Pose, Gaze, Annotations, Tactile Data	✓	×	✓	×	×
RFM-HRI (Ours)	**Robot Failures (and Recovery) in Medical HRI**	**FACS/AUs, Pose, Gaze, Transcripts, Survey Data**	✓	✓	✓	✓	✓

primarily observational rather than a physically interactive, time-critical collaboration setting [10]. The REACT [8] database contributes two data sets that capture user natural nonverbal reactions during human-robot interactions over time (e.g., a collaborative game and a photography scenario) and includes explicit evaluative feedback signals alongside nonverbal features and contextual traces [8]. REFLEX [19] extends the "reaction to failure" framing by capturing human reactions not only to induced robotic action failures but also to different explanation levels and strategies across repeated interactions, with dataset organization and modalities described in dedicated dataset sections [19]. Finally, HRI-SENSE provides a multimodal time-synchronized data set of social/physical behaviors and self-assessed questionnaires (e.g., frustration, satisfaction) during collaborative manipulation with a TIAGo robot, emphasizing multimodal signals and subjective impressions in embodied interactions [16].

2.4.4 Healthcare-Focused Multimodal Datasets. Finally, multimodal datasets focused on healthcare care, such as OpenRoboCare, demonstrate the value of rich sensing for care. OpenRoboCare provides expert occupational therapist demonstrations of Activities of Daily Living and spans five modalities (RGB-D video, pose tracking, eye-gaze tracking,

tactile sensing, and task/action annotations) to study physical caregiving routines [22]. The dataset is designed around expert demonstrations and caregiving principles rather than systematically induced robot failures and observed user reactions under time pressure [22].

Taken together, these resources motivate the need for datasets that jointly capture structured failures and multimodal human responses in high-stakes embodied tasks. Relative to prior work, our dataset is distinguished by (i) a healthcare-motivated *crash cart robot* guiding item retrieval, (ii) *structured, labeled failure injection* spanning speech/timing/comprehension/search, and (iii) aligned multimodal social signals *plus* post-hoc reflection after participants re-watch their own interaction [35].

3 User Study: Medical Human-Robot Interaction Failures

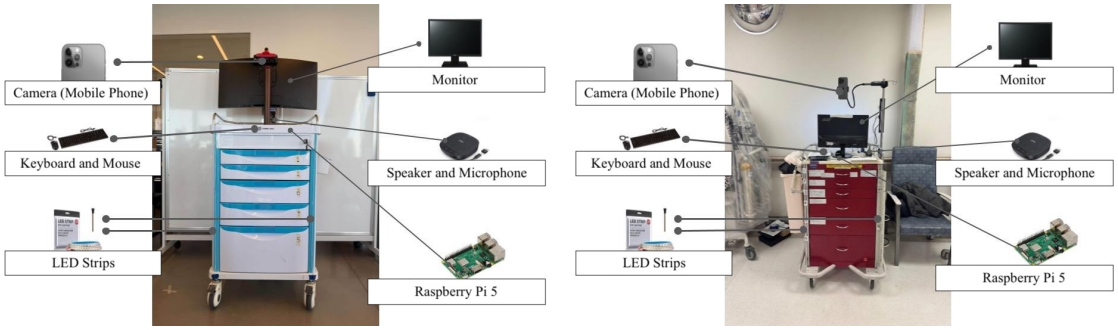
3.1 Task Overview, Robotic Platform, and Participants

3.1.1 Task Overview: Crash Cart Item Retrieval. In emergency and acute-care settings, clinicians frequently rely on crash carts to access medications and equipment during time-critical procedures [35]. Locating supplies on a cart can involve distributed coordination and decision-making under time pressure, as clinicians navigate equipment organization and communicate amid uncertainty [7]. Breakdowns in this process, such as delays in locating required items or miscommunication, have been linked to poorer procedural coordination and are especially consequential when rapid response is needed, motivating careful organization and rapid access to supplies [18]. To study how assistive robotic systems might support this form of situated, high-stakes collaboration, we focus on the task of guided item retrieval from a crash cart, using it as a representative interaction scenario in which timing, clarity of guidance, and shared understanding are essential.

3.1.2 Robotic Platform. In our work, we employ a semi-autonomous crash cart robot designed to support item retrieval tasks in healthcare settings, serving as the primary apparatus for the *RFM-HRI* dataset collection. Figure 2 showcases the assembled carts, both in our university and hospital setup. The platforms are built by retrofitting a standard crash cart to operate as a robot with multimodal cues, aligned with prior work in high-stakes healthcare settings, enabling realistic placement and interaction in clinical environments. To preserve validity during real-world hospital deployment, the drawers and shelves were stocked with standard medical supplies; however, for the controlled laboratory evaluation, we utilized distinct proxy items (e.g., office and hardware supplies) to simulate retrieval tasks. Table 2 details the specific inventory comparisons between the hospital and laboratory configurations. Visual cues are provided by addressable LED strips mounted alongside the drawer columns. Each column is associated with and identified by an LED segment, allowing the robot to highlight spatial locations using light cues. The robot also includes a verbal interface that uses a text-to-speech (TTS) engine to deliver standard prompts, status updates, and guidance messages. User speech is captured through an onboard microphone and later transcribed manually or via automatic speech recognition (ASR) with correction for downstream analysis.

The robot additionally incorporated an egocentric camera to capture first-person video of the interaction. Due to the distributed nature of the study, we did not enforce standardized camera hardware or video resolutions; however, the visual features reported in the dataset are standardized during post-processing (see Section 4.2.2).

To support post-hoc annotation and analysis, the robot maintained internal system logs indicating the start and end of each trial, as well as the specific scripted communication failure type (speech, timing, search, or comprehension) induced on each trial. While we only induced technical (system-level) failures, during interaction, the robot can experience



In Lab Data Collection **In Field Data Collection**

Fig. 2. Robotic crash cart platforms in real-world hospital deployment (left) and controlled university campus laboratory evaluation (right).

Table 2. Drawer-by-drawer inventory comparison of crash cart contents in the lab prototype vs hospital reference cart, supporting multimodal data collection and failure-induction studies.

Drawer #	Hospital Data Collection Supplies	In-Lab Data Collection Supplies
Drawer 1	Phenylephrine HCI, Epinephrine Injection, Metoprolol, Naloxone, Nitroglycerin, Norepinephrine, Adenosine	Blue tape, white tape, SD cards, LEDs
Drawer 2	Sodium Bicarbonate Injection USP, Lidocaine, 50% Dextrose Injection, 10% Calcium Chloride Injection	Orange pencils, yellow pencils, batteries, measuring tape
Drawer 3	Razor, Medical Tape, Surgical Gloves, Line Draw Arterial Blood Sample Syringe	Red pencils, wires, blue markers, blue pencils
Drawer 4	Blanket, Scissors, Face Masks	White board cleaner, circuit board, calculator, gloves, playstation controller, mouse
Drawer 5	1 mL Tuberculin Syringe, 6 mL Syringe, 500 mL IV Bag	Scissors, pliers, book, wipes, orange multimeter, hand sanitizer
Drawer 6	Drawer 6 is Empty in the Hospital Cart	No Drawer 6 on In-Lab Cart

failures from both technical (system-level) and experiential (user-level) perspectives. We define these failure types in Section 2.1.1.

3.1.3 Participants. We recruited 41 participants for an IRB-approved study through university campus advertisements and targeted outreach to Emergency Medicine listservs of Registered Nurses and medical residents in the Global North using convenience and purposive sampling. Of the 41 participants, 29 participated in lab studies and 12 in hospital studies. All participants completed a 30-minute study session and were compensated with a $10 gift card. Demographic information was collected voluntarily. 39% provided complete responses. Among those who responded, the sample included 11 male and 5 female participants, with ages ranging from 22 to 44 years (M = 25.8, SD = 6.4). Participants consisted of two groups, including 12 healthcare professionals and 29 non-experts. Participants in healthcare included Registered Nurses and Residents. Among healthcare workers, years of clinical experience ranged from 10 to 15 years. Lay participants were recruited from the general university population and reported their educational attainment, including Master's and doctoral/professional degrees. Participants were not required to have prior experience with robotic systems.

3.2 Experimental Methodology

3.2.1 Study Design. We conducted a within-subjects study with five conditions to examine how participants respond to robot failures during a crash cart item retrieval task. Each participant experienced all five conditions in randomized order, including four failure conditions (speech failure, timing failure, comprehension failure, and search failure) and one success condition (error-free interaction). Study sessions were scheduled for approximately 20 minutes, with first 7-8 minutes dedicated to interaction trials with the crash cart robot. Due to time constraints, a small number of participants completed slightly fewer or more than five trials; this introduces minor imbalance across conditions, which is accounted for in our analyses.

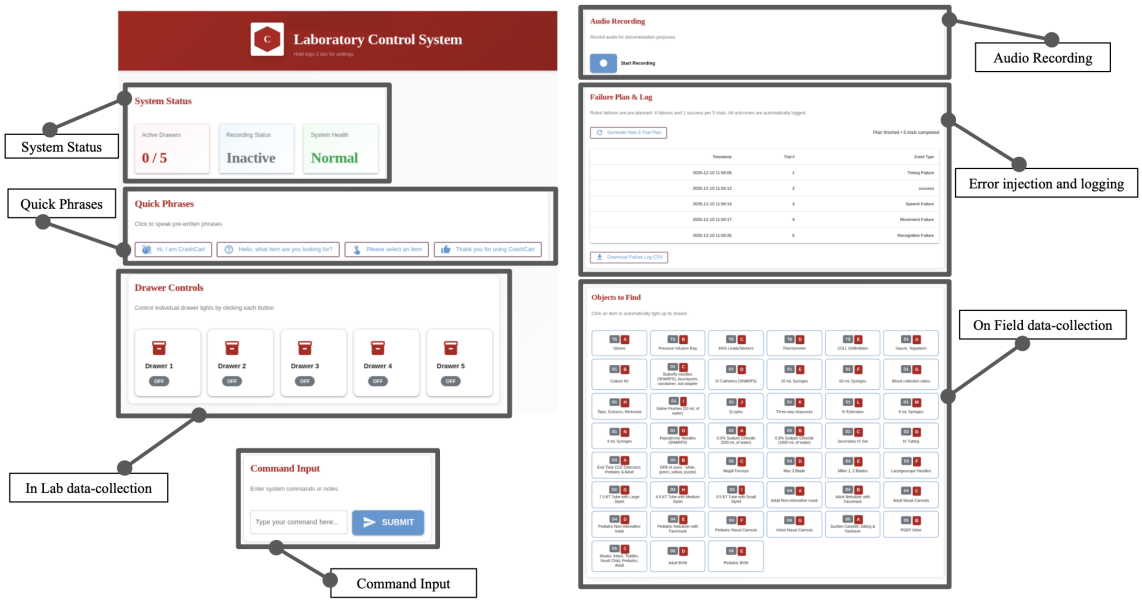
Participants completed a predetermined sequence of five trials, four of which included injected robot failures and one of which was error-free, presented in randomized order. This design allowed us to compare participant behavior and subjective experience across different types of failures while controlling for individual differences.

The study centered on a guided search and retrieval scenario in which the robot provided spoken instructions and visual clues to direct participants to specific items in the crash cart drawers. Failures were introduced systematically during interaction trials, (see Section 3.2.4). Participants were instructed to follow the robot's instructions as naturally as possible and were not informed in advance that failures would occur.

3.2.2 Study Setting. We conducted data collection in both hospital and laboratory settings to balance ecological validity with practical constraints on participant recruitment. In-situ studies with clinical professionals, particularly in emergency care contexts, are inherently limited by time pressure and extended shifts, making sustained access to emergency room nurses infeasible. To complement hospital-based data and enable broader participation, we therefore replicated the core retrieval task in a controlled laboratory environment. The laboratory setting featured a crash cart robot stocked with proxy items (Table 2), allowing systematic study of item retrieval behaviors with non-clinician participants while preserving task structure and interaction dynamics. In the hospital setting, the robot was deployed in a clinical training area and stocked with standard medical supplies (Table 2) to maintain realism. Together, this dual-setting design supports analysis across realistic clinical contexts and scalable laboratory conditions while accounting for real-world constraints on clinical participation.

3.2.3 Wizard-of-Oz Interface and Control. To facilitate real-time control during the study, we developed a custom web-based Graphical User Interface (GUI), shown in Figure 3a, which served as the primary control surface for the Wizard-of-Oz operator. This interface enabled the trained WoZ operator to manage multimodal robot behaviors, including addressable LED activation (Figures A.1c and A.1d), text-to-speech generation via pre-set "Quick Phrases" (Appendix Figure A.1a) custom text input (Appendix Figure A.1b), and audio data logging (Appendix Figure A.1b).

The interface organization was adapted to accommodate the differing inventory complexity between settings. In the laboratory evaluation, where the item set was limited, the operator utilized an external physical reference table to map requested items to their corresponding spatial locations, manually activating the specific drawer controls shown in Figure A.1c (e.g., toggling "Drawer 1"). Conversely, for the on-field hospital data collection, the high volume of medical supplies required a more streamlined approach. We therefore implemented a dedicated "Objects to Find" module within the GUI (Appendix Figure A.1d), mapping specific medical items (e.g., "EKG Leads", "Gloves") to their hardware addresses. This allowed the operator to select the requested object directly from a digital grid, automatically triggering the correct drawer and LED cues without the latency of manual lookup. Finally, the interface included the failure



(a) **Interface Overview:** The modular GUI features areas for monitoring, speech, and inventory.

Fig. 3. The Wizard-of-Oz (WoZ) Interface components (Part 1 of 4).

injection controls (Figure 4a), allowing the operator to trigger the system anomalies defined in Section 2.1.1 according to the study protocol.

Using this interface, we conducted a controlled Wizard-of-Oz (WoZ) study to investigate HRI dynamics during a clinically inspired object–retrieval task. Throughout all sessions, participants were led to believe that the robot operated autonomously. The WoZ operator was stationed in close proximity to the robot, positioned directly behind or adjacent to the cart, and controlled the system using the platform's onboard monitor, mouse, and keyboard. To maintain the illusion of autonomy despite this proximity, the operator's presence was framed as supervisory (e.g., monitoring system logs) rather than active control. Although the system currently relies on human-in-the-loop control, the WoZ protocol was designed to mirror the behavioral logic of a future autonomous crash cart assistant. The protocol follows best practices in Wizard-of-Oz HRI research and aligns with recent healthcare-focused robotic field deployments [34, 35].

Concretely, the WoZ operator executed all robot behaviors during the retrieval task using a standardized interaction script. Each interaction began with a standardized greeting by the WoZ: "Please retrieve X item". The participants responded verbally by asking the cart for the desired item. The WoZ operator then selected the corresponding item from a predefined inventory list, which was mapped internally to the physical storage location of the item. Upon selection, the cart generated a multimodal response: it verbally announced the target drawer or shelf and simultaneously illuminated the LED segment corresponding to the correct drawer. Items stored on the open top shelf were indicated through speech alone. Once the participant retrieved the item, the operator deactivated the visual cues and optionally activated a closing message (e.g., "Thanks for using the Crash Cart") to clearly signal task completion. At no point were

(a) **Failure Injection Log:** Sequences failures and logs trial outcomes.

Fig. 4. The Wizard-of-Oz (WoZ) Interface components (Part 2 of 4).

the failure conditions revealed to participants; if a participant questioned whether the robot was malfunctioning, the facilitator responded using a neutral, scripted reply to avoid biasing interpretation.

3.2.4 Failure Injection. We developed a dedicated failure injection module within the Wizard-of-Oz interface (previously described in Section 3.2.3) that triggers controlled breakdowns in robot behavior in terms of system perspective errors (see Section 2.1.1) during the drawer retrieval task. The module introduces four failure types that reflect common sources of miscoordination in clinical HRI. Each failure is executed through a function during live interaction.

The system failures are further explained below:

- *Speech Failure.* The robot provides an under-specified or incomplete instruction, by telling the participant to "Open the drawer." without specifying which drawer to open. For context, there are five drawers on the carts used for this study. Hence, this error captures ambiguity in robot communication.
- *Timing Failure.* The system introduces a deliberate delay of 3 secs (selected to disrupt the flow without causing total disengagement) before responding, followed by directing the human participant to the right drawer. This failure models disruptions in conversational timing and response latency.
- *Search Failure.* The robot identifies and highlights the wrong drawer through both speech and LED cues. This simulates guidance errors that occur when robots mislocalize objects or misinterpret the task state, or even mishear the item mentioned by the participant.
- *Comprehension Failure.* The robot signals that it cannot understand the participant's request. This represents breakdowns in speech understanding, which are common in noisy or dynamic environments such as the fast-paced emergency rooms. For example, after a participant requests a specific item (e.g., "epinephrine syringe"), the robot responds that it did not understand and asks for repetition, without offering any guidance.

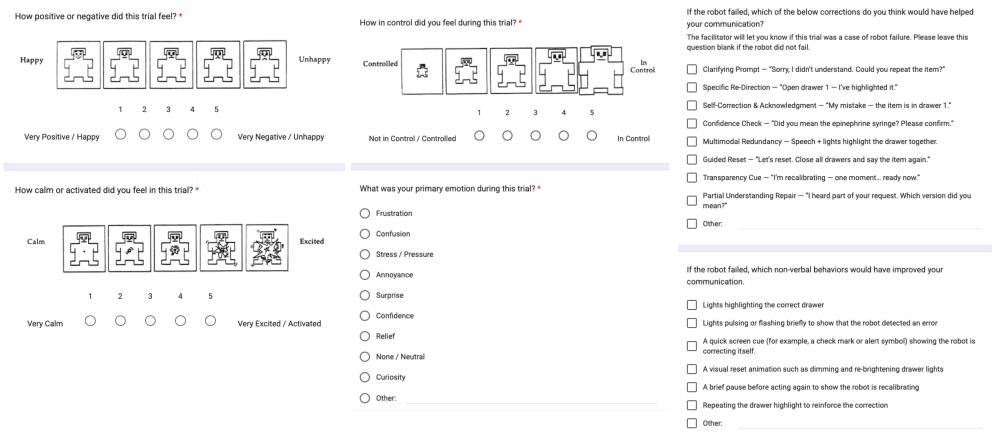


Fig. 5. Post-study survey form used to collect affective responses, including categorical emotion selection and Self-Assessment Manikin (SAM) ratings, as well as participants' preferred recovery strategies following each trial.

3.3 Study Procedure and Data Collection

Each study session consisted of four phases: an introduction, a guided object–retrieval block, a post-study reflection survey, and an optional demographic and background survey.

3.3.1 Introduction. After providing informed consent, participants were introduced to a crash cart robot designed to support item retrieval in a clinical setting. The facilitator explained that the robot's role was to help users locate medical supplies efficiently by providing spoken instructions and visual cues, and that the participant's task was to follow the robot's direction to retrieve requested items as accurately and quickly as possible. The participants were not informed that the robot could behave incorrectly or that failures would be introduced.

3.3.2 Guided Object–Retrieval Block. The core task consisted of a sequence of guided item-retrieval trials with the crash cart robot. Each participant completed a predetermined sequence of item retrieval trials with the crash cart robot. See Section 3.2.1 for details on the retrieval tasks and conditions. For each trial, the facilitator requested the participant to retrieve one item, and the robot responded with spoken instructions and visual cues to guide participants to the appropriate location.

3.3.3 Post-Study Reflection. We administered a survey to assess participants emotional reaction and recovery preferences. The full survey instruments are shown in Figure 5.

- *Emotional reaction.* For each trial, participants provided two forms of affective self-report. First, they selected their dominant emotional reaction, choosing from frustration, confusion, stress/pressure, annoyance, surprise, confidence, relief, curiosity, neutrality, or an open-ended option. Second, they reported their affective state using the Self-Assessment Manikin (SAM), which measures valence, arousal, and perceived control [5].
- *Recovery preference.* Participants also indicated which of several candidate recovery strategies they would have preferred the robot to take for the observed trial (e.g., offering an explicit apology and clarification, providing more step-by-step guidance, or escalating to a human teammate).

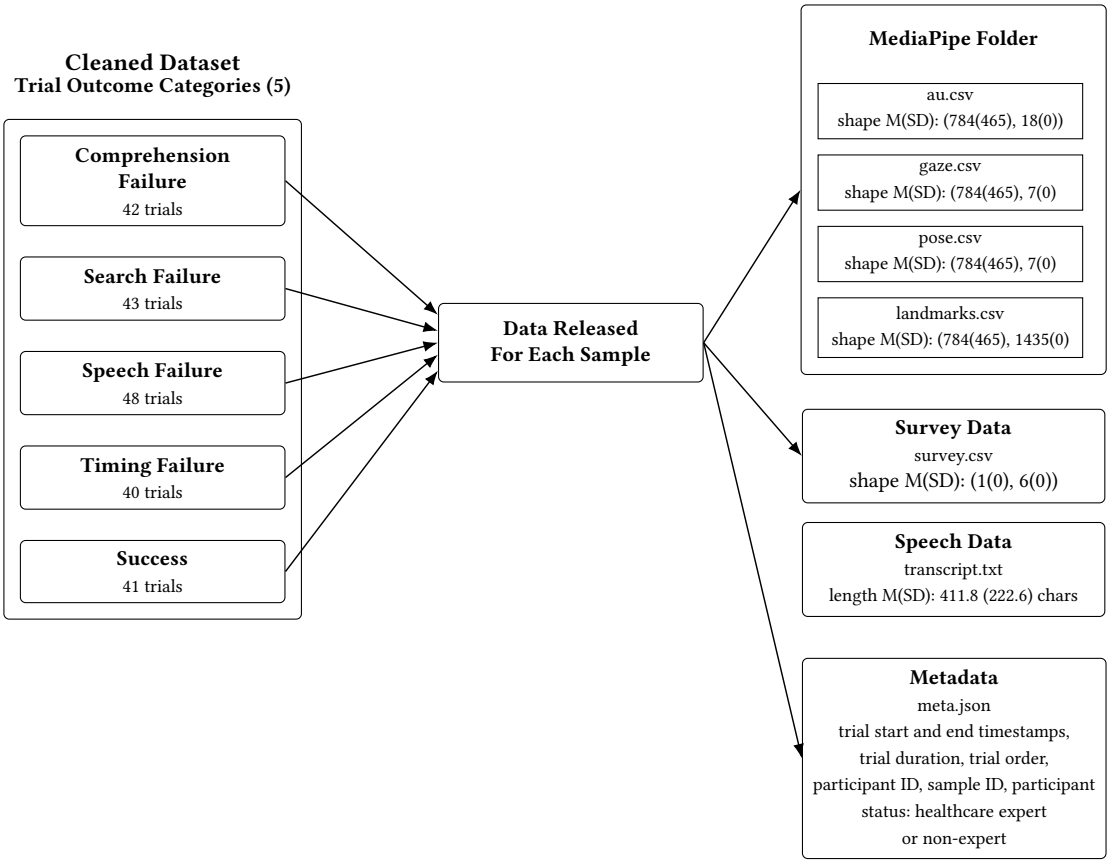


Fig. 6. **Dataset organization and per-trial multimodal composition.** The cleaned dataset is organized into five trial outcome categories (four injected robot failure types and success). For each trial, we release a self-contained sample comprising time-indexed visual behavioral features (MediaPipe-derived facial action units, gaze, pose, and landmarks), structured post-trial survey responses, and a natural-language task transcript. An overview of the dataset structure and accessible modalities is illustrated in Fig. 6. Together, these views summarize both the dataset-level structure and the multimodal signals available for modeling trial outcomes and recovery preferences, while abstracting away file-level organization.

3.3.4 Demographic Survey. Participants were asked to complete a brief, optional demographic questionnaire to support descriptive analysis of the study sample. The survey collected basic background information, including age range, gender identity, prior experience interacting with robots or automated systems, and familiarity with clinical or hospital environments. When time permitted, the demographic survey was administered immediately following the post-study reflection phase. In cases where session time was constrained, the same questionnaire was distributed to participants via follow-up email. All demographic responses were collected separately from interaction data and were used solely for aggregate analysis.

4 Dataset

4.1 Annotation Protocol and Labels

4.1.1 System-Level Failure and Trial Boundary Annotation. System-level annotations capture both (i) trial boundaries and (ii) robot communication failure types. Labels were assigned by a team of five trained researchers through manual review of recorded session videos, informed by the robot's internal logs and observable robot outputs. Each trial was annotated by a single researcher.

Trial segmentation was performed using consistent criteria across sessions. Trial start was defined as the onset of the facilitator's verbal request for the next item. Trial end was defined as the final spoken utterance before the subsequent trial began, or the end of the recording if no subsequent trial occurred.

Failure annotations indicate whether the robot behaved as designed or exhibited one of four scripted communication failures: speech, timing, search, or comprehension. Each failure type was associated with characteristic robot behaviors and utterances that served as cues during annotation. Speech failures involved under-specified instructions (e.g., "Open the drawer") or the absence of speech; timing failures involved delayed responses with apologetic cues (e.g., "Sorry, I am delayed…"); search failures involved incorrect spatial guidance (e.g., "I think the item is in drawer __"); and comprehension failures involved explicit non-understanding (e.g., "I did not understand that request."). System logs, verbal outputs, and visual context were jointly used to disambiguate failure types.

4.1.2 Annotation Workflow and Consistency. Annotations were produced by trained members of the research team following a shared, written annotation protocol that defined trial boundaries and failure categories (see Section 4.1.1). Annotators reviewed recorded session videos and referenced both system logs and observable robot behavior when assigning labels. Consistency was supported through the use of common definitions for trial start and end, explicit mappings between robot behaviors and failure types, and discussion-based adjudication of ambiguous cases. Participant-provided annotations were collected separately during the post-study reflection phase, as described below.

4.2 Data Processing and Feature Extraction

In our annotations, we focus *system errors*; we do not annotate or release user-error labels. We transformed the raw recordings into a set of aligned multimodal feature streams suitable for training and benchmarking machine learning algorithms. At a high level, the dataset includes four types of signals per trial. These include structured trial-level metadata, facial and head pose features, transcribed speech features, and survey responses. Detailed descriptions of these signals, and how they were extracted, can be found in the sections below.

4.2.1 Trial Segmentation and Metadata. Each participant session was segmented into trial-level interactions, where each trial corresponds to a single item retrieval attempt from initiation to completion. Trial segmentation was performed manually by trained researchers through review of recorded session videos, and clear definitions of what constituted trial start and trial end (see Section 4.1.1). Using these boundaries, we recorded structured trial-level metadata including failure type, trial start and end timestamps, participant identifier, trial index, sample identifier, and whether the participant was a healthcare expert or non-expert.

4.2.2 Facial and Head Pose Features. We analyze frame-level visual descriptors captured from the egocentric video, reflecting participants' facial behavior, head orientation, and visual attention during interaction with the crash cart robot. Specifically, we extract all visual features using Google MediaPipe's face analysis pipelines [24], which provide robust real-time estimation of facial landmarks and head pose from monocular RGB video.

For each session, we process the video stream with MediaPipe Face Mesh; this tool detects a dense set of 468 3D facial landmarks per frame, which use normalized image-space coordinates and relative depth. These landmarks encode fine-grained facial geometry across regions such as the eyebrows, eyes, nose, lips, and jawline. Using the detected landmarks, we estimate a 3D rigid head pose for the participant at each frame, defining it by rotation angles (pitch, yaw, roll) and translation with respect to the camera. We additionally derive coarse gaze-related features from eye-region landmarks, capturing attention shifts and visual search behavior during interaction breakdowns. In parallel, we extract frame-level facial action unit (AU) intensities using MediaPipe's Face Landmarker model with blendshape outputs. We map relevant blendshape coefficients to a subset of OpenFace-style action units [4] and scale their magnitudes to produce continuous AU intensity estimates per frame.

For downstream modeling and dataset release, we retain only frames with successful face detection and store the resulting facial landmark, head-pose, gaze-related, and facial action unit features as synchronized per-frame time series indexed by video frame number. Due to the distributed nature of the data collection effort, we did not enforce standardized cameras or video resolutions across recording sessions. Instead, we standardized the extracted behavioral signals during post-processing by operating on normalized facial landmarks and by producing temporally aligned and intensity-normalized facial action unit features, enabling consistent downstream analysis across heterogeneous video inputs.

4.2.3 Transcribed Speech Features. To capture speech data while preserving participant privacy, we converted the audio data of each session into time-stamped transcripts using the Whisper Speech Recognition model [29]. Whisper utilizes a Transformer-based encoder–decoder architecture, which researchers trained on hundreds of thousands of hours of multilingual speech. We use the model in English transcription mode to obtain segment-level transcripts with start and end timestamps. We retain the resulting time-stamped text segments for downstream analysis of participant–robot interaction dynamics. To balance privacy with the ability to model temporal dynamics in user reactions to robot failures, we release derived text features and time-stamped transcripts, but not the raw audio recordings.

4.2.4 Survey-Based Behavioral Signals. We include structured survey responses capturing participants' subjective experience of robot behavior on a per-trial basis, collected during the post-study reflection phase (see Section 3.3). For each trial, we release two categories of survey-derived signals. Participants' *emotional reactions* are represented using a categorical label indicating the dominant reported emotion and continuous affective ratings along the valence, arousal, and perceived control dimensions obtained using the SAM scale [5]. Participants' *recovery preferences* are represented as categorical labels corresponding to the single recovery strategy selected for each trial. All survey responses are released as structured, trial-level annotations aligned with the corresponding trial metadata. The full survey instruments are shown in Figure 5.

5 Data Analysis

We analyzed the dataset to characterize participant reactions to robot failures during object retrieval, and quantify how self-reported affect, discrete emotions, and recovery preferences vary across failure types, and success vs. failure outcomes.

5.1 Dataset Exploration and Quality Control

We performed a dataset-wide exploratory audit to characterize composition, balance, and the repeated-measures structure prior to inferential testing. We computed condition-wise distributions of the number of participants and

samples, including expert/non-expert breakdowns, the frequency distribution of reported emotions, and recovery preference selections stratified by modality (verbal, nonverbal, multimodal, and no recovery needed), providing a descriptive baseline for subsequent analyses.

We examined all (N=214) samples to assess the availability and integrity of the expected multimodal components (metadata, survey responses, speech transcripts, and visual behavioral features). For each modality, we checked schema-level consistency (e.g., expected fields/columns and data dimensional structure), identified and resolved issues such as missing or malformed data, and harmonizing representations where needed, and subsequently verified that the resulting data supported comparable analyses.

To characterize the repeated-measures structure, we quantified the samples-per-participant distribution and verified participant-to-sample mappings, to document the non-independent observation structure, which guided selection of statistical methods accounting for within-participant correlation in subsequent analyses. These checks were also used to confirm deviations from the nominal study design: 3 participants experienced fewer than 5 conditions and 4 participants experienced more than five conditions. Because participants reflected on these as unified experiences in their post-study responses, we excluded samples with multiple labels (conditions) from condition-based analyses to avoid ambiguous condition assignment, while retaining them in the released dataset.

We also quantified variation in survey instrument versions: samples from earlier participants (samples 1–102) included only verbal recovery strategy options, whereas later participants (samples 103–214) received an expanded survey with verbal, nonverbal, and multimodal options of recovery preferences. This variation is accounted for in our recovery preference analyses by reporting results separately by survey version and avoiding direct comparisons between strategies that appeared on different instruments. All subsequent analyses were conducted on the dataset after integrity verification and harmonization, with residual inconsistencies handled via predefined exclusions or carried forward as analysis constraints.

5.2 Primary Emotion Analysis

We analyzed self-reported *primary emotions* to characterize affective responses to robot performance during object retrieval. Only trials with a single condition label were included in inferential analyses. The analysis addressed four questions: (i) how primary emotions are distributed overall, (ii) whether emotion distributions differ by outcome (Success vs. Failure), (iii) whether they differ across condition types, and (iv) whether emotions change over time (trial order).

5.2.1 Descriptive Characterization. We first computed overall frequencies and percentages of primary emotions across all trials, followed by stratified summaries by outcome and by condition type. This step provided an interpretable baseline, identified dominant and sparse emotion categories, and guided which emotions were suitable for downstream modeling.

5.2.2 Emotions by Outcome and Conditions. We performed inspection of the stratified distributions that revealed clear variation in emotion prevalence across both outcome and condition types. To evaluate whether these observed differences reflected systematic effects rather than random variation, we tested for **associations** between primary emotion and (i) outcome (Success vs. Failure) and (ii) condition type (five levels: four failures and success). In both analyses, the independent variable (IV) was categorical (outcome or condition), and the dependent variable (DV) was the categorical primary-emotion label. The contingency-based analyses assume categorical variables and independent observations. While the measurement assumptions are satisfied by design, independence is violated due to repeated

measures within participants. We therefore interpret chi-square statistics as global measures of association and rely on within-participant permutation tests for valid significance estimation. Association strength is quantified using Cramér's V, which is emphasized over significance alone to characterize the magnitude of distributional differences.

5.2.3 Accounting for Repeated Measures. Because each participant contributed to multiple trials, trial-level observations were not independent. To obtain valid significance estimates under this repeated-measures structure, we computed permutation-based p-values by shuffling outcome or condition labels *within participants*. This preserves individual emotion response profiles while testing whether the alignment between task conditions and emotions exceeds chance expectations under clustering.

5.2.4 Temporal Analysis. To examine whether emotional responses evolved over time, we analyzed emotion patterns as a function of trial index. Trial-wise emotion frequencies were first inspected descriptively. For inferential quantification, we modeled the most frequent emotions as binary outcomes (presence vs. absence of a target emotion) using participant-clustered logistic models. Specifically, we estimated population-average effects with **generalized estimating equations (GEE)**, using trial index and outcome as predictors and clustering on participant. GEE provides robust inference under misspecification of the working correlation, making it well-suited for the present repeated-measures, sparse-category setting. GEE was selected because likelihood-based mixed-effects logistic models were unstable in this dataset due to sparse emotion categories and quasi-separation, leading to convergence failures and unreliable variance estimates. Emotions with insufficient occurrences were excluded from emotion-specific models.

Summary. Overall, descriptive analyses characterize the emotional landscape of interaction outcomes, contingency-based association tests validate global distributional differences across outcome and condition type, and participant-aware logistic models quantify how specific emotions vary with task progression over time.

5.3 Affective responses analysis

The affective analysis addressed three primary objectives:

(1) To determine whether affective responses differ between successful and failed interactions.
(2) To assess whether affective responses vary across the five interaction conditions (Success and four failure types).
(3) To evaluate the potential influence of trial order on affective responses.

5.3.1 Data Structure. Affective responses consisted of three 5-point ordinal variables (valence, arousal, and perceived control), recorded once per sample. We analyzed a subset restricted to one condition per sample (see Section 5.1). Given the ordinal, bounded nature of SAM ratings and the presence of within-participant dependence, all inferential procedures were selected to avoid assumptions of normality and sample-level independence.

5.3.2 Descriptive Characterization. As an initial step, we computed descriptive summaries for valence, arousal, and perceived control across the full dataset and stratified by interaction outcome and condition. Medians and interquartile ranges were used as primary descriptors, supplemented by means and standard deviations for completeness. This descriptive analysis provided an overview of affective response patterns and informed subsequent inferential test selection.

5.3.3 Affective Responses by Outcome. Examination of outcome-stratified affective rating distributions revealed clear and consistent shifts between successful and failed interactions. This strong descriptive structure motivated formal

testing of whether these differences represent systematic within-participant effects. For inferential analysis, each affective dimension was analyzed separately. The independent variable was interaction outcome (Success vs. Failure), and the dependent variable was the corresponding affective rating.

To respect the repeated-measures structure (several samples/trials per participant), affective ratings were aggregated at the participant level. For each participant, the median affective rating across success samples and the median affective rating across failure samples were computed, yielding one paired observation per participant. Paired differences were evaluated using the Wilcoxon signed-rank test, a non-parametric test that does not assume normality. Effect sizes were quantified using the matched-pairs rank-biserial correlation, providing an interpretable measure of the magnitude and direction of within-participant differences.

5.3.4 Affective Responses by Conditions. We examined condition-stratified distributions of affective ratings to characterize how valence, arousal, and perceived control vary across the five interaction conditions. These descriptive patterns were subsequently assessed using participant-level inferential analyses to evaluate condition-related differences in affective responses within participants. Each affective dimension (dependent variable) was analyzed separately, with interaction condition as a five-level categorical independent variable. Because each participant experienced multiple conditions, analyses were conducted within a repeated-measures (within-participant) framework. Condition-level inference was restricted to participants who contributed at least one valid observation for all five conditions. For participants with multiple trials within a condition, the median affective rating for that condition was used. Differences across conditions were assessed using the Friedman test (the non-parametric analogue of repeated-measures ANOVA). We quantified effect sizes using Kendall's W, capturing the degree of systematic variation in affective responses across conditions at the participant level.

5.3.5 Handling of Repeated Measures and Model Assumptions. Across all analyses, repeated measures (multiplicity of samples/trials per participant) were handled either by aggregating affective ratings within participant prior to testing or by using explicitly repeated-measures non-parametric tests. We avoided trial-level tests assuming independent observations. Linear mixed-effects models were explored as a sensitivity analysis but were not retained for inference due to identifiability issues arising from the limited and unbalanced condition–trial structure. Accordingly, rank-based within-participant methods were adopted as the primary inferential approach.

5.3.6 Temporal Considerations. Because trials were completed sequentially, affective responses may also reflect time-on-task or interaction fatigue effects independent of condition. To examine this possibility, we analyzed affective ratings descriptively as a function of trial index. Trial order was not treated as a primary inferential factor but was used to contextualize condition effects, as descriptive inspection did not reveal consistent or observable changes in affective ratings across trials.

Summary. In summary, we analyzed affective responses using a participant-aware, non-parametric framework that respects the ordinal nature of SAM ratings, accounts for **within-subject dependence**, and separates analytical procedures from reported results. All statistical outcomes and condition-specific patterns are presented in Section 6.3

5.4 Recovery Strategy Analysis

We analyzed participants' reported recovery strategy selections to characterize the recovery design space represented in the dataset and to document how recovery behaviors were distributed across communication modalities and strategy types. This analysis is descriptive in nature and focuses on dataset composition and structure. While modality-level

summaries were computed on restricted, comparable subsets (see Section 5.1), recovery strategy selections overall varied in number across participants; accordingly, we do not draw inferential claims about recovery preferences or optimal strategies at this stage.

5.4.1 Strategy Identification and Categorization. Recovery responses were parsed by separating comma-delimited selections into individual strategy mentions and normalizing textual variants. Across the dataset, this process yielded a set of distinct recovery strategies, including predefined survey options and participant-provided write-in responses, resulting in a total of 34 unique recovery strategy types, including strategies derived from expanded free-text responses. These strategies were retained as distinct labels for counting and reporting purposes. Each strategy was further categorized along two complementary dimensions. First, strategies were grouped by communication modality into *verbal*, *nonverbal*, and *multimodal* categories, where multimodal strategies explicitly combined speech with visual or embodied cues (e.g., simultaneous verbal explanation and visual highlighting). Second, verbal strategies were grouped by communicative function (e.g., clarification request, acknowledgment of error, explicit redirection, transparency cue).

5.4.2 Descriptive Summaries and Normalization. We computed frequency counts of recovery strategy mentions at multiple levels of aggregation, including per-strategy counts, per-category counts, and per-modality counts. Because participants could select multiple recovery strategies within a single trial, all summaries are reported in terms of mentions rather than trials. For modality-level comparisons, counts were normalized within the eligible subset of failure trials to report proportions that reflect relative composition rather than absolute sample size.

6 Results

In this section, we present preliminary results from an exploratory analysis of the RFM-HRI dataset. These analyses are intended to illustrate the range of affective, behavioral, and recovery-related signals captured in the data, rather than to draw causal conclusions. We summarized the dataset's composition and then examined user responses to robot failures during crash cart item retrieval, focusing on affective reactions across failure types and interaction outcomes. Finally, we reported initial patterns in user preferences for robot recovery behaviors, highlighting signals for failure acknowledgment and guidance. Together, these exploratory results operationalize the core contributions of this work: a multimodal, healthcare-inspired HRI dataset, an empirical characterization of user responses to robot failures, and early design-relevant insights into preferred recovery strategies. While preliminary, these findings underscore the utility of RFM-HRI for research on failure detection, user state modeling, and recovery policy design in healthcare and broader workplace robotics contexts.

6.1 Descriptive Statistics

The dataset contains **214 samples**: 64 of which were collected in the hospital with healthcare workers and 136 samples in laboratory settings with non-experts. For each sample, metadata files contain 6 keys capturing trial- and participant-level identifiers. Survey data consist of 6 questions per trial and cover **13 distinct emotion categories in total** and **34 unique recovery preference types** across all conditions. Visual features from image frame counts per trial had a mean of 784 (SD = 465). Verbal interaction transcripts had an average length of 411.8 characters per trial (SD = 222.6). An overview of dataset composition across conditions is provided in Table 3.

Table 3. Dataset overview across entire dataset (214 samples)

Condition	Participants	# Samples	All emotions (count)	Recovery preferences (count)
Comprehension	38 Total (11 Expert, 27 Non-Expert)	42 (13 Expert, 29 Non-Expert)	Annoyance (11) Confusion (11) Frustration (10) Curiosity (4) Surprise (3) Confidence (2) Relief (1)	Verbal (111) Nonverbal (1) Multimodal (17) No recovery needed (0)
Search	39 Total (12 Expert, 27 Non-Expert)	43 (15 Expert, 28 Non-Expert)	Confusion (16) Annoyance (14) Frustration (5) Stress / pressure (3) Confidence (2) Curiosity (2) None / neutral (1)	Verbal (110) Nonverbal (2) Multimodal (17) No recovery needed (0)
Speech	40 Total (11 Expert, 29 Non-Expert)	48 (15 Expert, 33 Non-Expert)	Confusion (17) Annoyance (9) Frustration (6) Confidence (5) Surprise (4) Relief (2) Stress / pressure (2) Amused (1) Curiosity (1) None / neutral (1)	Verbal (104) Nonverbal (5) Multimodal (36) No recovery needed (1)
Success	37 Total (10 Expert, 27 Non-Expert)	41 (13 Expert, 28 Non-Expert)	Surprise (12) Relief (10) Confidence (6) Confusion (4) Annoyance (3) Frustration (2) Curiosity (1) Happy (1) In doubt (1) None / neutral (1)	Verbal (23) Nonverbal (0) Multimodal (7) No recovery needed (2)
Timing	36 Total (10 Expert, 26 Non-Expert)	40 (12 Expert, 28 Non-Expert)	Frustration (12) Confusion (11) Annoyance (10) Curiosity (2) Surprise (2) Interest (1) Confidence (1) Relief (1)	Verbal (113) Nonverbal (2) Multimodal (13) No recovery needed (0)

6.2 Primary Emotions

6.2.1 Analysis Set. From 214 total samples, 25 trials contained multiple condition labels (e.g., trials where two failures occurred consecutively) and were excluded from condition-based inferential tests. The final analysis set comprised $N = 189$ trials from 41 participants (trials per participant: 1–5, $M = 4.6$). Of these, 153 were failure trials and 36 were success trials.

6.2.2 Descriptive Characterization. Across all analyzed trials ($N = 189$), the most frequent emotions were Confusion ($N = 49$, 25.9%), Annoyance ($N = 42$, 22.2%), and Frustration ($N = 33$, 17.5%), followed by Surprise ($N = 19$, 10.1%), Confidence ($N = 16$, 8.5%), Relief ($N = 12$, 6.3%), Curiosity ($N = 8$, 4.2%), and Stress/Pressure ($N = 5$, 2.6%). Five categories were sparse ($N < 5$ each: Amused, None/Neutral, In Doubt, Happy, Interest) (Table 4).

Table 4. Overall distribution of primary emotions across analyzed samples ($N = 189$).

Emotion	Count	Percentage (%)
Confusion	49	25.9
Annoyance	42	22.2
Frustration	33	17.5
Surprise	19	10.1
Confidence	16	8.5
Relief	12	6.3
Curiosity	8	4.2
Stress / Pressure	5	2.6
Amused	1	0.5
None / Neutral	1	0.5
In Doubt	1	0.5
Happy	1	0.5
Interest	1	0.5
Total	**189**	**100.0**

6.2.3 Emotions by Outcome. Emotion distributions differed substantially by outcome (Success vs. Failure). Failure trials were dominated by negative cognitive-evaluative states: Confusion (47/153, 30.7%), Annoyance (40/153, 26.1%), and Frustration (31/153, 19.1%). In contrast, success trials most often elicited Surprise (11/36, 30.6%), Relief (9/36, 25.0%), and Confidence (6/36, 16.7%) (Figure 7). A chi-square test on the full 2×13 contingency table showed a significant association between outcome and emotion, $\chi^2(12) = 77.40$, $p < .001$, with a large effect size (Cramér's $V = 0.640$). Because trials are clustered within participants, we additionally computed participant-aware permutation p-values by shuffling outcome labels *within* each participant (10,000 permutations), yielding a permutation $p < .0001$ (reported as $p = 0.0000$ with 10,000 permutations). Figure 8 visualizes these outcome-level differences.

6.2.4 Emotion by Condition. Emotion distributions also varied across the five conditions (Success, Speech Failure, Timing Failure, Comprehension Failure, Search Failure). The 5×13 chi-square test indicated a significant association between condition type and emotion, $\chi^2(48) = 112.76$, $p < .001$, with a medium effect size (Cramér's $V = 0.386$). A within-participant permutation test (10,000 permutations; shuffling condition labels within participant) again yielded $p < .0001$. These confirm that the observed association was unlikely to arise from participant-specific response patterns or violations of sample independence.

Condition-wise counts are shown in Figure 9. Confusion was most frequent for Speech Failure ($N = 15$) and for Search Failure ($N = 14$), while Frustration was especially prominent for Timing Failure ($N = 12$). Success trials were characterized by high Surprise ($N = 11$) and Relief ($N = 9$), with comparatively low Confusion ($N = 2$) and Annoyance ($N = 2$).

Temporal Analysis. Trial-wise emotion frequencies revealed systematic changes across the five trial indices (1–5). Confusion was most prevalent early in the session (45.0% in Trial 1) and declined across subsequent trials, whereas Frustration increased over time, peaking in Trial 4 (28.2%). Annoyance showed a moderate increase through Trial 4 (30.8%) followed by a decrease in the final trial (17.1%). In contrast, emotions associated with resolution or positive appraisal (e.g., Surprise, Relief, Confidence) exhibited relatively stable frequencies across trials without pronounced monotonic trends (Figure 10). Participant-clustered GEE logistic models confirmed these temporal patterns for the most prevalent emotions: trial index was associated with a significant decrease in the odds of reporting Confusion

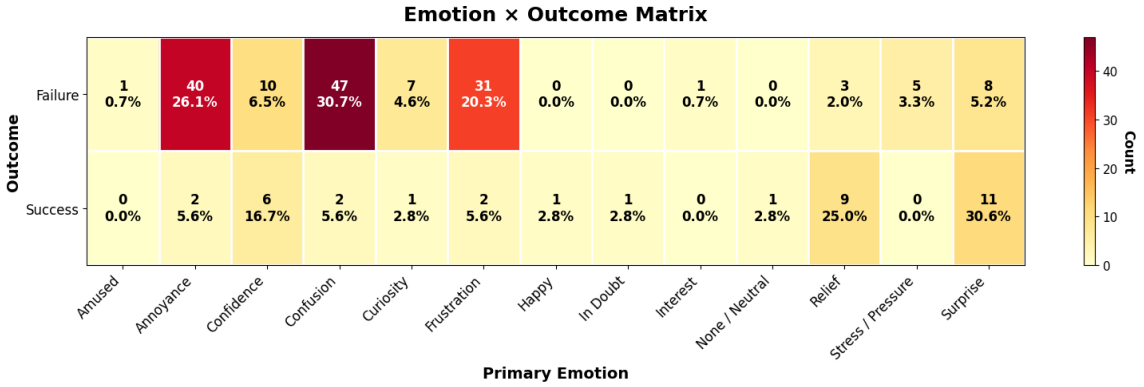


Fig. 7. Primary emotion distributions by outcome (Success vs. Failure)

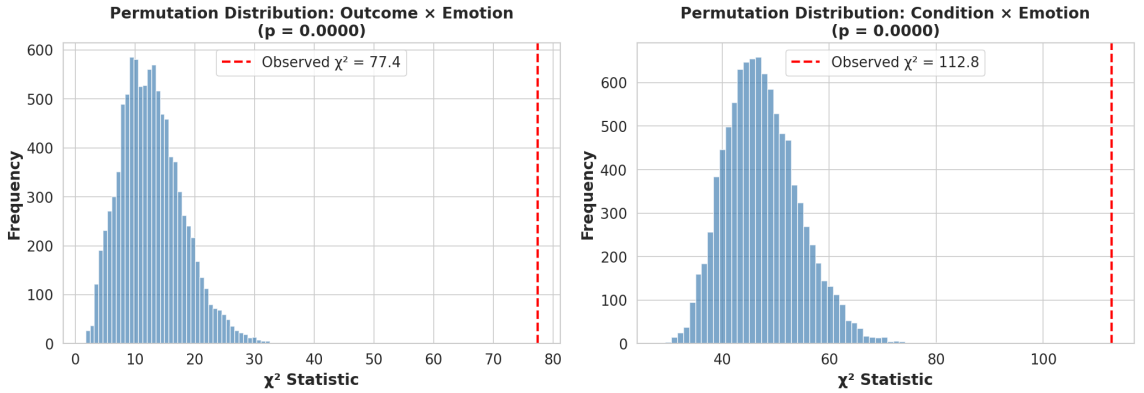


Fig. 8. Permutation distributions for emotion – outcome (left) and emotion – condition (right) associations.

($\beta = -0.289, p = 0.021$) and a significant increase in the odds of reporting Frustration ($\beta = 0.288, p = 0.017$), while the effect for Annoyance was not significant ($\beta = 0.118, p = 0.31$).

6.3 Affective Response Analysis (SAM Ratings)

Analysis Set. Participants reported affective responses after each trial using the Self-Assessment Manikin (SAM), providing ratings along three dimensions: *Valence* (unpleasant–pleasant), *Arousal* (calm–excited), and *Control* (low–high sense of control). Ratings were collected on 5-point ordinal scales. The affect dataset comprised 189 trials from 41 participants. Twenty-five trials with multiple condition labels were excluded to ensure unambiguous condition attribution. No trials were excluded due to missing or invalid affect data.

Descriptive Characterization. Across all trials, median affective ratings indicated moderately negative experiences overall (Table 5). Valence ratings were centered toward the unpleasant end of the scale, while arousal and control ratings were moderate, suggesting emotionally engaged but not highly activated interactions.

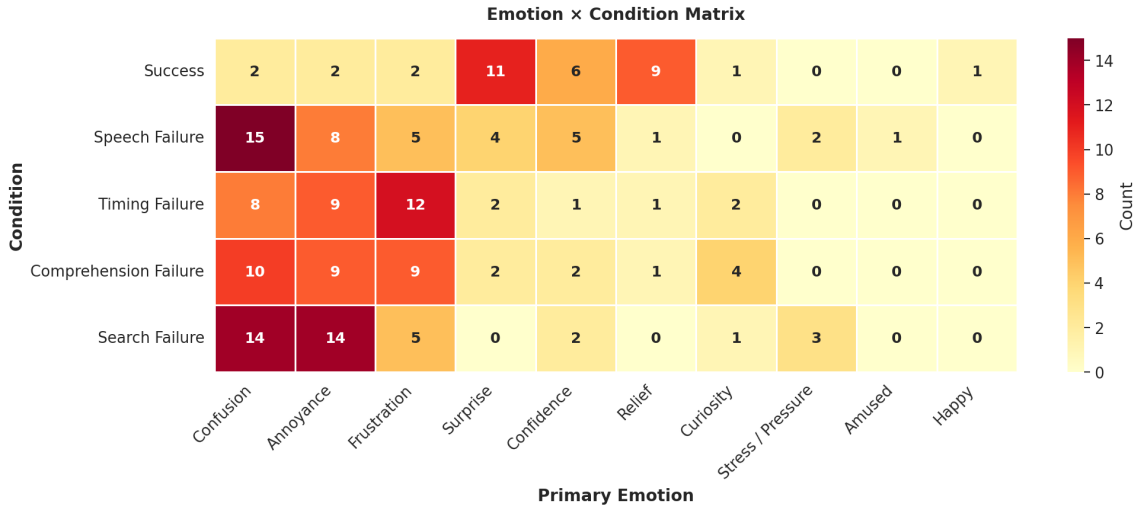


Fig. 9. Primary emotion distributions across interaction conditions

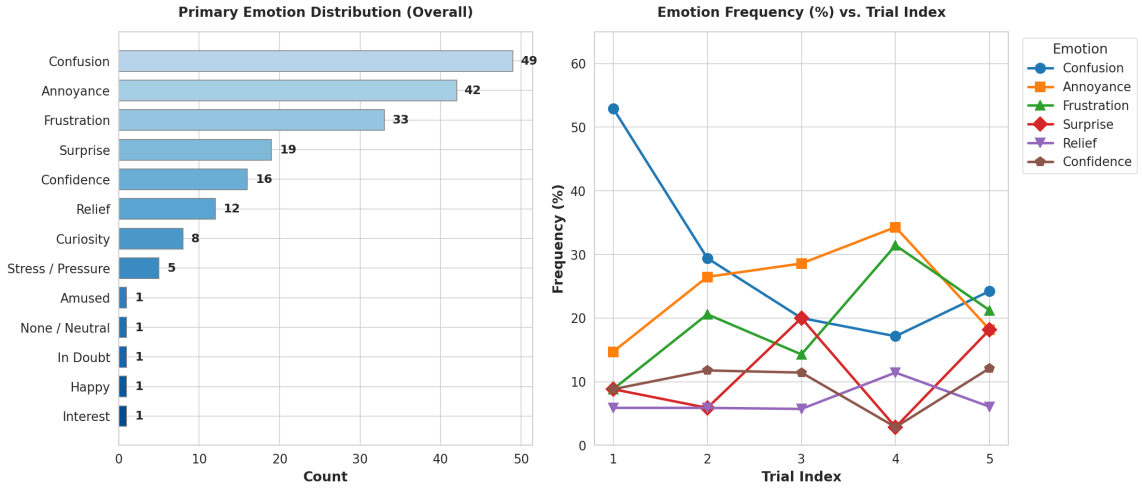


Fig. 10. Overall and temporal distributions of primary emotions

Table 5. Overall affective ratings across all trials (N = 189).

Dimension	Median	IQR	Mean	SD
Valence	4.0	[2.0–4.0]	3.28	1.38
Arousal	3.0	[2.0–4.0]	2.92	1.18
Control	3.0	[2.0–4.0]	2.83	1.33

Affective Responses by Outcome. we conducted analysis to 34 participants with data in both outcome categories (Success, Failure). Paired Wilcoxon signed-rank tests revealed robust outcome-level differences across all three affective

Table 6. Paired Wilcoxon comparisons of affective ratings between Success and Failure outcomes (participant-level summaries, $n = 34$).

Dimension	W	p-value	Rank-biserial r	Median (Succ vs Fail)
Valence	1.0	1.73×10^{-6}	0.996	1.0 vs 3.8
Arousal	81.0	0.0158	0.538	2.0 vs 3.0
Control	88.0	0.00165	−0.645	4.0 vs 2.5

Table 7. Friedman tests for affective differences across five interaction conditions ($n = 28$ participants with complete data).

Dimension	χ^2	df	p-value	Kendall's W
Valence	49.27	4	5.13×10^{-10}	0.440
Arousal	11.20	4	0.0244	0.100
Control	13.89	4	0.00765	0.124

Table 8. Median and interquartile range (IQR) of affective ratings by interaction condition.

Condition	n	Valence		Arousal		Control	
		Mdn	IQR	Mdn	IQR	Mdn	IQR
Comprehension Failure	37	4.0	[3.0–4.0]	3.0	[2.0–4.0]	2.0	[2.0–4.0]
Search Failure	39	4.0	[3.0–5.0]	3.0	[2.0–4.0]	2.0	[1.0–3.5]
Speech Failure	41	4.0	[3.0–4.0]	3.0	[2.0–4.0]	3.0	[2.0–4.0]
Timing Failure	36	4.0	[3.0–4.0]	3.0	[2.0–4.0]	3.0	[1.75–3.0]
Success	36	1.0	[1.0–2.0]	2.0	[1.0–4.0]	4.0	[3.0–5.0]

dimensions (Table 6). Successful interactions were associated with significantly more positive valence ($\text{median}_{\text{Succ}} = 1.0$ vs. $\text{median}_{\text{Fail}} = 3.8$, $W = 1.0$, $p = 1.73 \times 10^{-6}$, $r_{\text{rb}} = 0.996$), lower arousal ($\text{median}_{\text{Succ}} = 2.0$ vs. $\text{median}_{\text{Fail}} = 3.0$, $W = 81.0$, $p = 0.0158$, $r_{\text{rb}} = 0.538$), and higher perceived control ($\text{median}_{\text{Succ}} = 4.0$ vs. $\text{median}_{\text{Fail}} = 2.5$, $W = 88.0$, $p = 0.00165$, $r_{\text{rb}} = -0.645$) compared to failed interactions. Effect sizes indicated large effects for valence and control and a moderate effect for arousal.

6.3.1 Condition-Level Differences. We conducted Friedman tests on $N = 28$ participants with complete data for all conditions ($k = 5$). Significant differences were observed for valence, arousal, and control (Table 7). Kendall's W indicated a strong effect for valence and small-to-moderate effects for arousal and control. Descriptively, failure conditions were associated with higher valence ratings (more unpleasant experiences) and lower perceived control, whereas success trials were characterized by low valence ratings and high control (Figure 11). Median affective summaries by condition are reported in Table 8.

Correlational Analyses. Participant-level Spearman correlations (computed over median ratings across trials) revealed a moderate negative association between valence and control ($\rho = -0.436$) and between arousal and control ($\rho = -0.471$), while valence and arousal were largely uncorrelated ($\rho = -0.008$). These correlations suggest that more unpleasant or arousing interactions were associated with reduced perceived control.

Temporal Analysis. Figure 12 suggests no visible temporal variation.

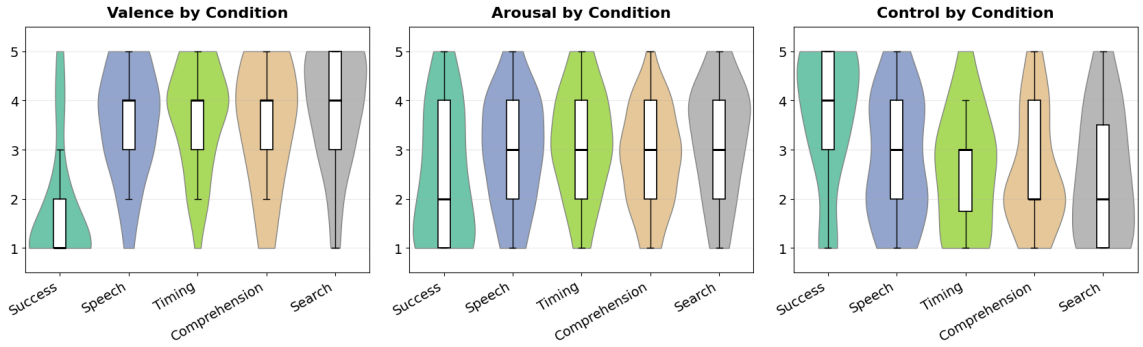


Fig. 11. Affective ratings distribution across interaction conditions

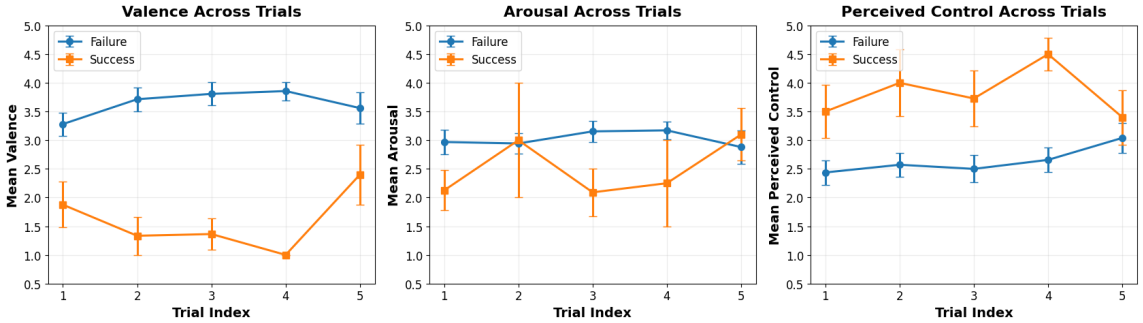


Fig. 12. Affective ratings across trial order

6.4 Recovery Preferences

All included failure samples contained both the verbal and nonverbal recovery questions (i.e., no missing survey versions or files), yielding a fully eligible subset for modality-level summaries.

6.4.1 Recovery Modality Distribution. Across the eligible subset, recovery preferences were dominated by verbal strategies (Table 9). Participants produced 431 verbal mentions (63.9% of mentions) and 219 nonverbal mentions (32.5%). Explicitly multimodal strategies (combined verbal + nonverbal in the same response) were rare (24 mentions; 3.6%), and no responses indicated that recovery was unnecessary (0 mentions).

6.4.2 Strategy diversity. Recovery preferences were heterogeneous, comprising 44 unique strategy strings overall, including 27 unique verbal strategies and 18 unique nonverbal strategies. The most frequently mentioned verbal strategies emphasized communicative clarity and transparency. Table 10 lists the ten most common verbal strategies, led by clarifying prompts (74 mentions) and self-correction/acknowledgment (59 mentions), followed by specific re-direction (45), transparency cues (42), and confidence checks (42).

Table 9. Recovery modality distribution (eligible subset; N = 153 failure trials). Percentages are computed over all recovery mentions.

Modality	Mentions	Percent (%)
Verbal	431	63.9
Nonverbal	219	32.5
Multimodal	24	3.6
None needed	0	0.0
Total	**674**	**100.0**

Table 10. Top verbal recovery strategies by number of mentions (eligible subset; N = 153 failure trials).

Verbal strategy (normalized label)	Mentions
Clarifying prompt (e.g., request repetition/clarification)	74
Self-correction & acknowledgment	59
Specific re-direction (explicit drawer/item instruction)	45
Transparency cue (state/explanation, e.g., recalibrating)	42
Confidence check (explicit confirmation)	42
Guided reset (restart/close drawers and retry)	41
Partial-understanding repair	36
Multimodal redundancy (speech paired with lights)	24
Other (rare; long-tail responses)	1

7 Discussion

7.1 Comparison to Prior Work

Our findings extend prior HRI failure research. First, while datasets like ERR@HRI 2.0 [9] and REFLEX [19] examine conversational errors, we demonstrated that embodied item retrieval tasks produce distinct patterns, including confusion (30.7%), frustration (19.1%), and reduced perceived control (M=2.61 vs. M=3.61, p<.001). The negative correlation between valence and control (ρ = -0.54, p < .001) indicates a deleterious effect of the loss of agency in robot failures. Also, existing social signal detection work [6, 32] recognizes failures but does not provide preferred recovery preferences for failure types. Our finding that 63.9% of users prefer verbal transparency challenges the assumption that embodied robots primarily need physical recovery cues. Recovery preferences did not vary significantly by failure type ($\chi^2(27) = 21.55$, p=.760), suggesting that universal principles (transparency, acknowledgment) outweigh strategies specific to failure types. While deployments [35] identify real world robot failures, we demonstrate these can be systematically reproduced in controlled settings. Participants reported confusion as the primary emotion in several first trials but noted that this decreased over time, while frustration increased across trials. This trend confirms that repeated failures shift users from uncertain to irritated, providing quantitative evidence for RF-HIP predictions [17].

7.2 Design Implications for Robot Failure Recovery

Our findings suggest several implications for robot designers. First, robots should prioritize full transparency about failures. Participants consistently preferred recovery strategies that combined explicit verbal correction with transparency cues. Specifically, verbal strategies such as specific re-direction ("I meant syringes are in drawer 4") and self-correction acknowledgments ("Sorry, I made a mistake"), alongside non-verbal signals like lights blinking on the correct drawers or visual reset animations. This preference for multimodal, transparent recovery suggests that users want robots to

clearly signal both *that* an error occurred and *how* to proceed. Second, recovery strategies that include both verbal acknowledgment and actionable guidance are more effective than apologies alone. Non-verbal recovery cues (such as LED lights or visual animations) can complement verbal recovery but may be insufficient on their own, particularly for failures that leave users uncertain about what went wrong. This finding aligns with conversational repair literature [2] but extends it to physically grounded interaction. Third, effective recovery may require matching the modality of the recovery strategy to the nature of the failure, though our data suggest this should be applied cautiously. Finally, user preferences may diverge from effective strategies in practice. That 63.9% of preferences emphasized verbal recovery, despite the embodied item-retrieval task, suggests that users may underestimate multimodal redundancy. Our dataset enables testing whether preferences align with actual task performance.

7.3 Limitations

Our study has several limitations that constrain the generalizability of our findings. First, we used a Wizard-of-Oz protocol to ensure consistent failure injection across participants. While this enabled systematic data collection, it means our findings reflect user responses to simulated rather than real world robot failures. Autonomous failure detection and recovery systems may produce different user reactions, particularly if users perceive errors as reflecting genuine technical limitations (e.g., in terms of hardware and software) rather than experimental simulations. Second, the crash cart robot was not deployed during real medical procedures, as the healthcare workers' interactions were collected alongside hospital corridors. Third, our dataset captures only single-session interactions. We cannot assess how user reactions, trust, or recovery strategy preferences evolve with repeated exposure to robot failures over long term interactions. Longitudinal studies are needed to understand whether users adapt to characteristic failure types, whether social cues remain reliable indicators of failure perception over time, and how trust calibration unfolds in clinical roles where emotional restraint is normative. Fourth, our sample of 41 participants, although including both healthcare workers and laypersons, limits our statistical power to detect interactions among failure type, user expertise, and recovery strategy preferences. Finally, our proposed failure modes focus on communication breakdowns (speech, timing, comprehension, search) and do not capture physical level errors such as cart navigation failures, collision avoidance errors, or integration failures with hospital information systems. These additional failure modes may elicit different user reactions and require different recovery strategies.

7.4 Future Directions

There are several potential directions from this work. First, **closing the loop from detection to recovery**: our dataset provides the multimodal signals needed to train classifiers that detect failure types from facial expressions, speech patterns, and behavioral cues, and to develop policies that select recovery strategies based on predicted failure type and user state. Because the dataset includes both objective behavioral traces and subjective recovery preferences, researchers can investigate whether user-preferred strategies align with objective efficacy, and can develop adaptive recovery systems that balance user preferences with task performance.

Second, **longitudinal trust calibration in clinical teams**: future work should examine how healthcare teams adapt to robots that make errors over extended deployments. Critical questions include whether repeated exposure to failures degrades or refines trust, how team coordination patterns shift when robots introduce uncertainty, and whether failure experiences lead to automation bias or appropriate reliance. Such studies would extend our single-session findings to understand the cumulative impact of robot failures on clinical workflow and team dynamics.

A third promising direction involves **autonomous failure management systems**: transitioning from Wizard-of-Oz simulation to fully autonomous robots that detect their own failures, infer user reactions, and execute context-appropriate recovery strategies. Our dataset provides a foundation for developing and benchmarking such systems, and future work involves validation of candidate recovery policies in real or high-fidelity clinical settings where failures have authentic consequences.

8 Conclusion

We presented RFM-HRI, a multimodal dataset that captures user reactions to systematically induce robot failures in a crash cart robot during an item retrieval task. Through Wizard-of-Oz studies with 41 participants across laboratory and hospital settings, we recorded synchronized facial expressions, head pose, speech, and self-reported responses to four failure types derived from three years of field observations. Our analysis revealed that confusion, annoyance, and frustration dominate user emotional responses to robot failures, with users consistently preferring recovery strategies that combine transparent verbal acknowledgment with actionable guidance. Recovery strategy preferences varied by failure type, suggesting that effective recovery may require matching recovery modalities to failure characteristics.

This work contributes to the HRI community by providing an empirically grounded set of failure modes for crash cart robots and robots created for item retrieval tasks, a publicly available multimodal benchmark dataset of robot failures with user reactions (RFM-HRI), and design insights for failure detection and recovery in physically grounded interactions. By capturing both the robot's technical perspective and the user's experiential perspective on failures, our dataset enables future research on adaptive recovery systems that can detect failures from multimodal behavioral signals and select context-appropriate recovery strategies. We hope this resource will advance the development of robots that not only acknowledge their limitations but actively work to repair interactions when failures inevitably occur.

References

[1] Henny Admoni. 2016. Nonverbal communication in socially assistive human-robot interaction. *AI Matters* 2, 4 (Dec. 2016), 9–10. doi:10.1145/3008665.3008669

[2] Essam Alghamdi, Martin Halvey, and Emma Nicol. 2024. System and User Strategies to Repair Conversational Breakdowns of Spoken Dialogue Systems: A Scoping Review. In *ACM Conversational User Interfaces 2024*. ACM, Luxembourg Luxembourg, 1–13. doi:10.1145/3640794.3665558

[3] Deepali Aneja, Daniel McDuff, and Mary Czerwinski. 2020. Conversational error analysis in human-agent interaction. In *Proceedings of the 20th ACM international conference on intelligent virtual agents*. 1–8.

[4] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE. https://github.com/TadasBaltrusaitis/OpenFace

[5] Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The Self-Assessment Manikin and the Semantic Differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1 (1994), 49–59. doi:10.1016/0005-7916(94)90063-9

[6] Alexandra Bremers, Alexandria Pabst, Maria Teresa Parreira, and Wendy Ju. 2024. Using Social Cues to Recognize Task Failures for HRI: Overview, State-of-the-Art, and Future Directions. doi:10.48550/arXiv.2301.11972 arXiv:2301.11972 [cs].

[7] FJR Calvo et al. 2025. Improving Good Practices for Patient Safety in an Emergency Department through Simulation and Resource Management. *Journal of Emergency Medicine Simulation and Practice* (2025). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12567365/ Simulation and cognitive resource management in ED.

[8] Kate Candon, Nicholas C Georgiou, Helen Zhou, Sidney Richardson, Qiping Zhang, Brian Scassellati, and Marynel Vázquez. 2024. REACT: Two datasets for analyzing both human reactions and evaluative feedback to robots over time. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 885–889.

[9] Shiye Cao, Maia Stiber, Amama Mahmood, Maria Teresa Parreira, Wendy Ju, Micol Spitale, Hatice Gunes, and Chien-Ming Huang. 2025. Err@ hri 2.0 challenge: Multimodal detection of errors and failures in human-robot conversations. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 14130–14135.

[10] Yuchen Cui, Qiping Zhang, Brad Knox, Alessandro Allievi, Peter Stone, and Scott Niekum. 2021. The empathic framework for task learning from implicit human feedback. In *Conference on Robot Learning*. PMLR, 604–626.

[11] Devleena Das, Siddhartha Banerjee, and Sonia Chernova. 2021. Explainable AI for Robot Failures: Generating Explanations that Improve User Assistance in Fault Recovery. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21)*. 351–360. doi:10.1145/3434073.3444657

[12] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2012. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia* 19, 3 (2012), 34–41.

[13] Dave Epstein, Boyuan Chen, and Carl Vondrick. 2020. Oops! predicting unintentional action in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 919–929.

[14] H.A. Frijns et al. 2024. Human-in-the-loop error detection in an object manipulation robotic system. *Frontiers in Robotics and AI* (2024). doi:10.3389/frobt.2024.1356827

[15] Manuel Giuliani, Nicole Mirnig, Gerald Stollnberger, Susanne Stadler, Roland Buchner, and Manfred Tscheligi. [n. d.]. Frontiers | Systematic analysis of video data from different human–robot interaction studies: a categorization of social signals during error situations. ([n. d.]). https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2015.00931/full

[16] Balint Gucsi, Nguyen Tan Viet Tuyen, Bing Chu, Danesh Tarapore, and Long Tran-Thanh. 2025. HRI-SENSE: A Multimodal Dataset on Social and Emotional Responses to Robot Behaviour. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 1319–1323.

[17] Shanee Honig and Tal Oron-Gilad. 2018. Understanding and Resolving Failures in Human-Robot Interaction: Literature Review and Model Development. *Frontiers in Psychology* 9 (June 2018). doi:10.3389/fpsyg.2018.00861 Publisher: Frontiers.

[18] GA Jacquet et al. 2018. Emergency Department Crash Cart Contents and Review. *Journal of Emergency Medicine Equipment* (2018). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5847507/ Crash cart organization and rapid access importance.

[19] Parag Khanna, Andreas Naoum, Elmira Yadollahi, Mårten Björkman, and Christian Smith. 2025. Reflex dataset: A multimodal dataset of human reactions to robot failures and explanations. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 1032–1036.

[20] Jean-Claude Laprie. 1995. Dependable Computing and Fault Tolerance: Concepts and Terminology. In *Proceedings of the 25th International Symposium on Fault-Tolerant Computing*. Pasadena, CA, USA, 2–11.

[21] Shan Li, Weihong Deng, and JunPing Du. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2852–2861.

[22] Yifan Liang et al. 2025. OpenRobocare: A Large-Scale Open Dataset for Human-Robot Interaction in Healthcare. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM. https://openrobocare.github.io/

[23] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, 94–101.

[24] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).

[25] Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. [n. d.]. Frontiers | To Err Is Robot: How Humans Assess and Act toward an Erroneous Social Robot. ([n. d.]). doi:10.3389/frobt.2017.00021

[26] Terran Mott, Cailyn Smith, Aaron Fanganello, and Tom Williams. [n. d.]. We Don't Ask Such Things: A Mixed-Methods Reassessment of Bounded Proportionality in Robotic Norm Violation Response. ([n. d.]).

[27] Huy Quyen Ngo, Elizabeth J. Carter, and Aaron Steinfeld. 2024. Human Perception of Robot Failure and Explanation During a Pick-and-Place Task. *Proceedings of the AAAI Symposium Series* 4, 1 (Nov. 2024), 373–379. doi:10.1609/aaaiss.v4i1.31819

[28] Maria Teresa Parreira, Sukruth Gowdru Lingaraju, Adolfo Ramirez-Aristizabal, Manaswi Saha, Michael Kuniavsky, and Wendy Ju. 2024. A Study on Domain Generalization for Failure Detection through Human Reactions in HRI. doi:10.48550/arXiv.2403.06315 arXiv:2403.06315 [cs].

[29] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv preprint arXiv:2212.04356* (2022). https://arxiv.org/abs/2212.04356

[30] Micol Spitale, Maria Teresa Parreira, Maia Stiber, Minja Axelsson, Neval Kara, Garima Kankariya, Chien-Ming Huang, Malte Jung, Wendy Ju, and Hatice Gunes. 2024. Err@ hri 2024 challenge: Multimodal detection of errors and failures in human-robot interactions. In *Proceedings of the 26th International Conference on Multimodal Interaction*. 652–656.

[31] Maia Stiber and Chien-Ming Huang. 2020. Not All Errors Are Created Equal: Exploring Human Responses to Robot Errors with Varying Severity. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*. ACM, Virtual Event Netherlands, 97–101. doi:10.1145/3395035.3425245

[32] Maia Stiber, Russell H Taylor, and Chien-Ming Huang. 2023. On using social signals to enable flexible error-aware hri. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 222–230.

[33] Tauhid Tanjim, Promise Ekpo, Huajie Cao, Jonathan St George, Kevin Ching, Hee Rin Lee, and Angelique Taylor. 2025. Human-Robot Teaming Field Deployments: A Comparison Between Verbal and Non-verbal Communication. doi:10.48550/arXiv.2506.08890 arXiv:2506.08890 [cs].

[34] Tauhid Tanjim, Jonathan St_George, Kevin Ching, and Angelique Taylor. 2025. Help or Hindrance: Understanding the Impact of Robot Communication in Action Teams. (2025).

[35] Angelique Taylor, Tauhid Tanjim, Michael Joseph Sack, Maia Hirsch, Kexin Cheng, Kevin Ching, Jonathan St.G̃eorge, Thijs Roumen, Malte F. Jung, and Hee Rin Lee. 2025. Rapidly Built Medical Crash Cart! Lessons Learned and Impacts on High-Stakes Team Collaboration in the Emergency Room. *arXiv preprint arXiv:2502.18688* (2025). https://arxiv.org/abs/2502.18688

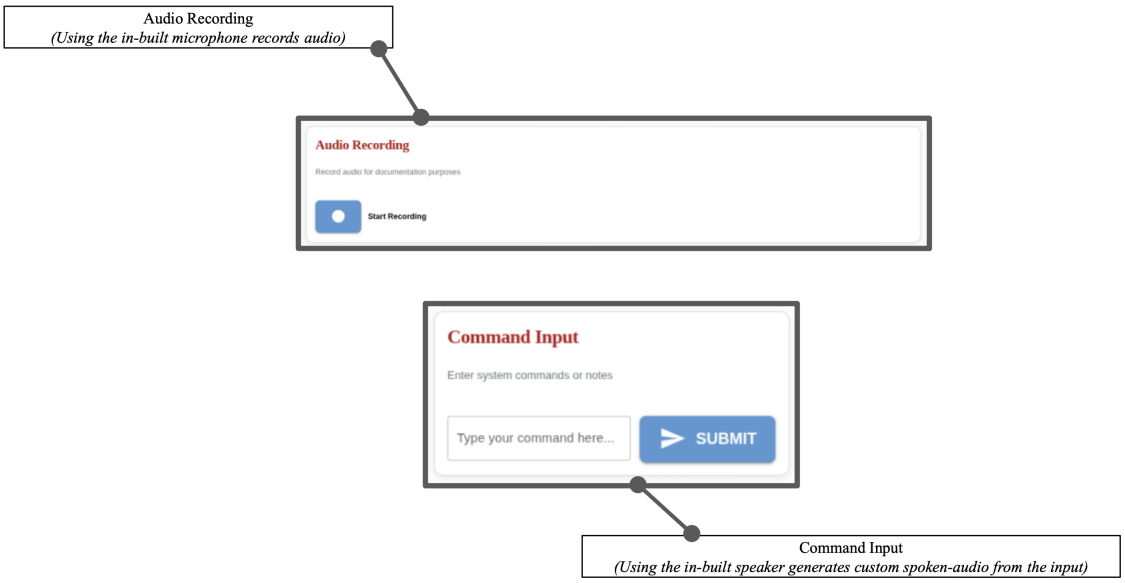
[36] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. 2013. A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 1–6.

[37] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. 2011. Facial expression recognition from near-infrared videos. *Image and vision computing* 29, 9 (2011), 607–619.

A Wizard-of-Oz Interface Details

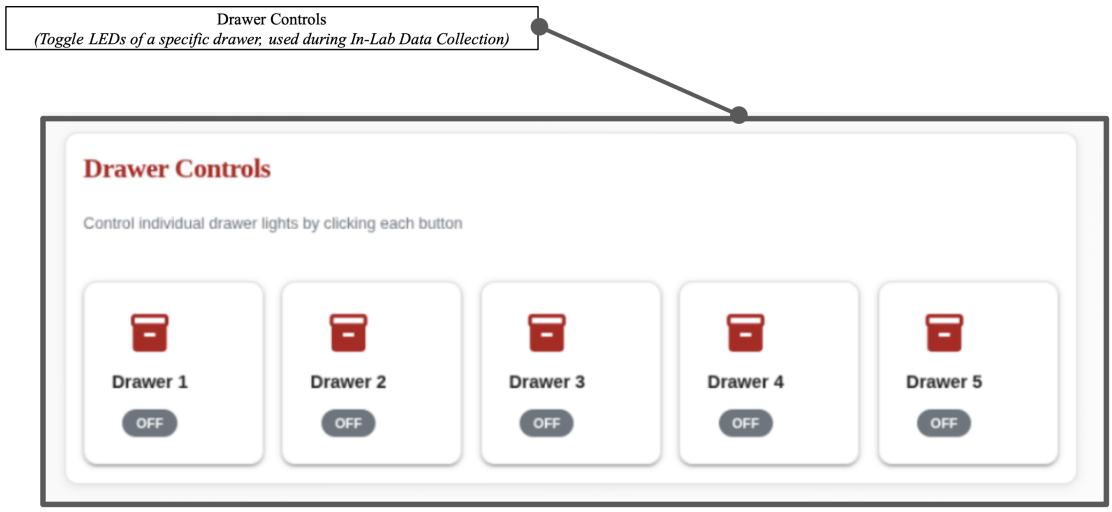
In this appendix, we provide the complete visual breakdown of the Wizard-of-Oz interface used in both laboratory and hospital settings.

(a) **Status & Speech:** Monitors active drawers/health. "Quick Phrases" trigger pre-set TTS.

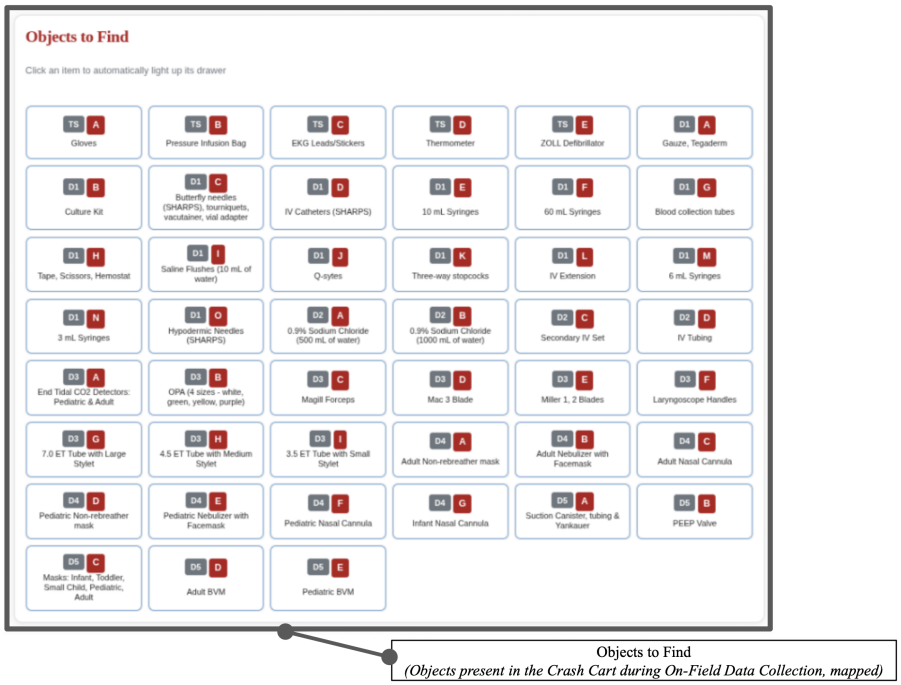


(b) **Audio & Input:** Recording controls and text input for custom TTS responses.

Fig. A.1. The Wizard-of-Oz (WoZ) Interface components (Part 3 of 4).



(c) **Lab Mode (Manual):** The "Drawer Controls" module used for the in-lab proxy task.



(d) **Hospital Mode (Mapped):** The "Objects to Find" grid used for the on-field medical task.

Fig. A.1. The Wizard-of-Oz (WoZ) Interface components (Part 4 of 4).