frontiers | Frontiers in Virtual Reality

TYPE Original Research
PUBLISHED 25 April 2025
DOI 10.3389/frvir.2025.1536393

OPEN ACCESS

EDITED BY
Anna Kobusinska,
Poznań University of Technology, Poland

REVIEWED BY
Asterios Leonidis,
University of Crete, Greece
Tomas Trescak,
Western Sydney University, Australia

*CORRESPONDENCE
Arthur Caetano,
caetano@ucsb.edu

RECEIVED 28 November 2024
ACCEPTED 31 March 2025
PUBLISHED 25 April 2025
CORRECTED 01 September 2025

CITATION
Caetano A, Aponte A and Sra M (2025) A design toolkit for task support with mixed reality and artificial intelligence.
Front. Virtual Real. 6:1536393.
doi: 10.3389/frvir.2025.1536393

A design toolkit for task support with mixed reality and artificial intelligence

Arthur Caetano[1]*, Alejandro Aponte[2] and Misha Sra[1,2]

[1]Department of Computer Science, University of California Santa Barbara, Santa Barbara, CA, United States, [2]Department of Media Arts and Technology, University of California Santa Barbara, Santa Barbara, CA, United States

Efficient performance and acquisition of physical skills, from sports techniques to surgical procedures, require instruction and feedback. In the absence of a human expert, Mixed Reality Intelligent Task Support (MixITS) can offer a promising alternative. These systems integrate Artificial Intelligence (AI) and Mixed Reality (MR) to provide realtime feedback and instruction as users practice and learn skills using physical tools and objects. However, designing MixITS systems presents challenges beyond engineering complexities. The complex interactions between users, AI, MR interfaces, and the physical environment create unique design obstacles. To address these challenges, we present MixITS-Kit—an interaction design toolkit derived from our analysis of MixITS prototypes developed by eight student teams during a 10-week-long graduate course. Our toolkit comprises design considerations, design patterns, and an interaction canvas. Our evaluation suggests that the toolkit can serve as a valuable resource for novice practitioners designing MixITS systems and researchers developing new tools for human-AI interaction design.

KEYWORDS

mixed reality, artificial intelligence, design toolkit, human-AI interaction, task assistance, prototyping

1 Introduction

Mixed reality (MR) technology can display interactive virtual content anchored to the real environment (Milgram et al., 1995) This capability is suitable to offer instructions to tasks situated in the real environment (Curtis et al., 1999). Such instructions can be contextualized to users' needs using artificial intelligence (AI), with examples including early instances of AI such as rule engines (Feiner et al., 1993) and more recent developments such as large language models (LLMs) (Wu et al., 2024). AI has also been used to offer personalized feedback in real world tasks through MR interfaces, leveraging sensing and error detection techniques (Anderson et al., 2013). In this work, we refer to the class of MR systems that display AI-generated instruction and feedback in a task situated in the real environment as Mixed Reality Intelligent Task Support (MixITS). MixITS systems have been proposed for several task domains including cooking (Sosnowski et al., 2023; Wu et al., 2024), machine maintenance (Feiner et al., 1993), and fitness training (Anderson et al., 2013; Mandic et al., 2023)

MixITS have the potential to improve user performance and facilitate physical skill acquisition by overcoming limitations of other forms of instruction and feedback. Textual and audiovisual formats, such as books and video-tutorials, lack personalization and proactive interventions to prevent or correct user errors, all of which can be achieved

with MixITS. Human coaches can benefit from the deployment of MixITS systems as they allow tracking and realtime analysis of multiple physiological and kinematic signals including full body pose, eye-gaze, and heart rate (Bernal et al., 2022) which could go unnoticed by humans. In the absence of a human expert due to schedule, budget or geographical constraints, MixITS can act as an automated coach to ensure consistent training and performance of end-users.

MixITS could enable a large audience to acquire new physical skills and perform tasks in the real world in a safe and precise manner. However, research and design of MixITS can be challenging due to the inherited complexity of AI and MR technologies. Previous research has identified challenges that AI introduces to HCI (Yang et al., 2020), emphasizing the need for new design guidelines and toolkits specific to AI-based applications. Separate studies have pointed factors of MR that contribute to complexity in design, such as understanding of users' surroundings and relationships between virtual and real elements (Freitas et al., 2020). Although the existing AI (Amershi et al., 2019; Yang et al., 2020; Yildirim et al., 2022; Feng et al., 2023; LaViola, 2017; Ultraleap, 2024; Meta, 2024; Apple, 2024; Microsoft, 2024; Council, 2024) provide a useful starting point, considerably less research has been done on design aids at the intersection of these technologies (Xu et al., 2023).

In addition to the challenges inherited from AI and MR, the application domain of task support situated in the real environment adds another layer of complexity. Examples of critical design decisions in MixITS include: How much guidance should be provided to balance correct execution and learning? Should the system always intervene in case of mistake, or should it balance the negative impact on the outcome and keep the user focused on the task? How to promote user trust on the system given the AI can make mistakes and provide instructions that diverge from the users' embodied knowledge of the task? Limited research has been done in eliciting MixITS design challenges and developing structured design tools in this domain are non-existent.

To fulfill this gap, we propose MixITS-Kit, a structured set of design tools to tackle the complexity of MixITS at multiple abstraction levels, contributing to faster developments in this application domain. Our toolkit is comprised of three elements:

1. Interaction Canvas: a visual tool to streamline the analysis of interactions between the users, system, and the environment (Section 5.4).
2. Design Considerations: a catalog of high-level design considerations in MixITS (Section 5).
3. Design Patterns: a set of lower level describing common problems in MixITS, observed solutions, and application examples to guide prototyping and development (Section 5.3).

We created MixITS-Kit based on observation and documentation of eight low-fidelity MixITS prototypes developed by 25 graduate students during a graduate 10-week human-AI interaction (HAI) class. We used Don Norman's Gulfs of Execution and Evaluation model as the foundation for our Interaction canvas that allows designers to analyze potential interaction problems between the actors involved in a MixITS scenario. The Interaction Canvas also enabled us to systematically identify problems and corresponding design solutions in the prototypes, facilitating design pattern elicitation (Winters and Mor, 2009; Retalis et al., 2006; Borchers, 2000). Through a reflexive thematic analysis (Braun and Clarke, 2006; Braun and Clarke, 2024) of the projects, we distilled six overarching design considerations for MixITS systems.

To evaluate the benefits and limitations of our toolkit, we conducted a user study where a separate group of nine participants used MixITS-Kit to solve a series of design problems. This evaluation assessed ease of use, toolkit research goals proposed in prior literature (Ledo et al., 2018), and potential use as shared vocabulary. Our results demonstrate that participants were able to utilize MixITS-Kit to propose multiple solutions for common challenges in MixITS design. User feedback also points to future work in an interactive version of our toolkit, especially to streamline gulf analysis and navigation in the design pattern catalog.

MixITS-Kit offers a structured set of tools that fulfill gaps in existing design aids that consider only AI or MR alone and tackle the unique challenges of task support situated in the real environment from different levels of abstraction. MixITS-Kit contributes to consolidating and disseminating design practices in MixITS design, favoring faster development of better systems that can help a broader audience to learn physical skills or perform real-world tasks in a safe and efficient manner.

2 Related work

In this section, we review literature on MR task support, which is similar to MixITS but lacks an AI component. We then focus on prior research in MixITS systems, which addresses technical challenges in underlying technologies but provides limited tools for interaction design. Finally, we discuss methods used in earlier work to derive design guidelines for human-AI interaction and mixed reality design. While these tools are useful for MixITS design, they overlook many unique problems of this domain, which we address in later sections.

2.1 Mixed reality task support

Supporting users in tasks situated in the real environment has been a long lasting motivation for MR research. A 1990 pioneer study demonstrated the feasibility of using MR to instruct workers in an aircraft wiring assembly task (Thomas and David, 1992; Curtis et al., 1999). However, evaluations showed no significant reduction in task completion time. Research associated this result with the need for improved instructional design, ergonomic issues and social acceptance of the equipment—concerns that remain relevant today.

Later studies compared MR instructions with alternatives in several real environment tasks. In a warehouse pick-up task, researchers demonstrated improvements in task performance when using MR instructions instead of paper and audio baselines (Weaver et al., 2010; Schwerdtfeger et al., 2006). In a manual assembly task, an MR instruction condition over performed paper and 2D displays in terms of error rate and mental effort (Tang et al., 2003). In a machinery repair task, MR instructions

allowed mechanics to locate tasks more quickly and with less head movement than when using a 2D display (Henderson and Feiner, 2010). ARDW designed and evaluated an MR system to support printed circuit board testing (Chatterjee et al., 2022) and reported reduced context-switching between instruction and the task, significantly faster localization on the board, and highlighted the need for fine alignment between virtual elements and physical elements.

Tiator et al. (2018) proposed an MR instruction system to assist climbers, highlighting benefits such as maintaining appropriate challenge levels and documenting ascents. In Skillab (Shahu et al., 2023), researchers combined MR instructions with haptic assistance through electrical muscle stimulation in a floor lamination task. Their system offered better user experience and outcome quality than a paper-based baseline. Their prototype identified the need to allow users to modify specific steps according to their preferences and prior knowledge. This work demonstrates the potential of combining MR instruction with physical actuators and the need for adaptability. MR has been used to support piano learning with positive effects on learners' motivation, ability to read music scores, and play notes correctly (Rigby et al., 2020). In the same study, researchers emphasize the need of offering customization options to meet users' preferences and skill level.

These relevant studies and systems offer evidence of the benefits of MR instruction for tasks situated in the real environment and contribute to design considerations that resonate with the ones derived from our own study. However, they do not include the AI component present in the MixITS definition. Our work corroborates previous findings, such as the need to adapt to users' preferences and skill level, and potential to incorporate physical actuators and additional sensors missing in current commercial MR devices. Differently from prior work, MixITS-Kit adds to their design considerations by accounting for the presence of AI in the design.

2.2 Mixed reality intelligent task support

More directly related to our work, a pioneer system proposed by Feiner et al. (1993) presented a augmented reality (AR) interface combined with a knowledge-base to automate the design of instructions for maintenance and repair tasks. This is one of the earliest examples of what we refer to in this paper as MixITS. Their work identified potential for multimodal instructions and suggested MixITS will become the preferable way to learn complex real environment tasks in the future. More recent work in MixITS include models of user-gaze behavior too estimate visual attention and expertise level in cooking and coffee brewing tasks (Yoo et al., 2023). ARTiST (Wu et al., 2024) used LLM few-shot prompting to optimize the length and the semantic content of textual instructions displayed in MR for two tasks—cooking and spatial localization. This intervention alleviated participants' cognitive load and improved task performance when compared with unprocessed instructions. Also in a cooking task, Rheault et al. (2024) proposed a MixITS that leverages action recognition, error and object detection to intelligently provide instruction and feedback.

To facilitate the design and evaluation of MixITS systems, prior work has also proposed technical platforms including open source MR systems combined with vision and language models (Bohus et al., 2024) and a visual analytics system (Castelo et al., 2023). Although these platforms alleviate the engineering and data analytics burden, they offer limited guidance to the user-centered design of novel MixITS systems. Our work aims to address this gap, providing a toolkit that is specifically designed for MixITS systems. By doing so, we hope to encourage the development of more diverse MixITS applications. Currently, very few such systems exist, limiting our ability to evaluate them at scale. As more MixITS systems are built and deployed, we will be better positioned to identify and understand new real-world challenges as they emerge, furthering our knowledge and improving future designs in this critical area of human-AI interaction.

2.3 Designing with AI

Researchers have developed human-AI interaction guidelines and explored the role of designers in AI teams through design workshops, product reviews, and literature analyses. These studies have also identified design challenges stemming from AI's uncertain capabilities and complex outputs.

For example, Amershi et al. (2019) developed 18 human-AI interaction design guidelines, derived from over 150 recommendations from academic and industry sources, and validated through evaluations including HCI practitioner testing. The authors noted the need for specialized tools to address unique design challenges in AI-integrated interfaces, particularly in modeling interactions involving the physical context. A follow-up study presents an evaluation protocol for human-AI interaction, focusing on productivity applications such as document and slide editors, search engines, email, and spreadsheet applications (Li et al., 2023).

Delving into AI system design, Yang et al. (2020) identified two key complexities that designers face: uncertainty about AI capabilities, and varying complexity of the types of output AI models generate. They proposed a four-level classification, with level-four systems being the most challenging to design due to continuous learning and adaptive outputs. MixITS systems fall into this category. The authors suggest that existing design methods are inadequate to address the range of potential AI behaviors in real-world contexts for these complex systems.

Exploring how experienced designers use AI in enterprise applications, Yildirim et al. (2022), found AI was used in both UI design and higher-level systems and services. They identified tools such as data-annotated service blueprints and wireframes that designers use to work with AI. The study highlights the need for AI-specific design tools, arguing that traditional UX practices require adaptation to address the unique challenges and opportunities presented by AI technologies.

Expanding on previous studies, Feng et al. (2023) conducted a contextual inquiry with 27 industry UX practitioners, revealing key challenges in AI integration. UX Practitioners encountered difficulties in collaborating with AI teams in user-centric design due to late involvement. Many struggled to communicate key AI performance metrics such as accuracy. Explainability was also a

concern, as unreliable AI behavior caused loss of user trust. To mitigate these issues, authors introduced "AI model fidelity" and "probabilistic user flows" to aid designers in AI application development. They also noted the importance of involving domain experts in collaborative human and AI design approaches, even if they lack UX or AI expertise. These tools are particularly relevant for MixITS systems, where AI behavior uncertainty significantly impacts users.

While tools and guidelines from prior work provide valuable insights for general AI interaction design, they overlook many of the unique MixITS challenges derived from the combination of MR and tasks situated in the real environment. For example, AI understanding of the physical world might be misaligned with the users' embodied knowledge, resulting in incorrect instructions. Another design problem not addressed by current HAI guidelines is timeless and adequate modality for AI interactions, that might ignore the real environment task at hand is priority for the user. Informed by prior HAI research and careful analysis of eight MixITS prototypes, our MixITS-Kit offers specialized design tools for intelligent task support in mixed reality.

2.4 Designing with MR

Challenges of designing with MR described in literature (Ashtari et al., 2020) include difficulty to anticipate users' movements in the real environment and dealing with distractions from the real environment. Other studies highlighted the difficulties of aligning physical and virtual elements spatially and semantically (Ellenberg et al., 2023). To better support MR designers, industry and academia have proposed design tools.

Industry practitioners have compiled valuable guidelines for designing MR experiences. Guidelines such as "design to avoid occlusion" and "design for the interaction zone" can help designers to create more usable and comfortable interactions in MR considering tracking limitations and user upper body reach limits (Ultraleap, 2024). Recommendations range from spatial layout to interaction such as "size and distance for proper depth perception" and "give users (. . .) multi-modal input, such as hand ray-and-speech input (. . .)" to help guide the design of MR experiences (Meta, 2024). Best practices suggest displaying content within the user's field of view, supporting "indirect gestures" (i.e., gestures executed in resting position), and avoiding the display of overwhelming motion to "prioritize comfort" (Apple 2024). Industry has also proposed design processes and techniques for designing MR applications. Low fidelity prototyping techniques such as "bodystorming," (Burns et al., 1994; Oulasvirta et al., 2003), i.e., manipulating low-cost, tangible props that represent components of a MR application, are a useful method to validate and refine early design concepts (Microsoft, 2024). Acting out a scenario is recommended to gain perspective on how a user would interact with an MR application (Microsoft, 2024). Additional guidelines for frequent tests and design iterations to better understand user behavior are also suggested (Council, 2024; Meta, 2024). We incorporated these techniques into the assignments for our design course to help students prototype MixITS systems without feeling constrained by current technological limitations or implementation challenges.

LaViola (2017), proposed practical guidelines for designing 3D UIs—including MR interfaces—such as ergonomics and comfort (e.g., "Design for comfortable poses"), user safety (e.g., "Provide physical and virtual barriers to keep the user and the equipment safe"), and interaction design (e.g., "Consider using props and passive feedback, particularly in highly specialized tasks"). A survey on human remote collaboration through MR reviewed extensive research and identified design choices in local interfaces that can be helpful in MixITS systems as well (Wang et al., 2021). The same study presents a comprehensive list of technological toolkits, but does not point to design toolkits. The domain of human remote collaboration holds similarities to MixITS but the substitution of the remote expert by and AI creates unique challenges, for example, in modulating instruction and feedback frequency and handling false positive error detection, motivating our research on a specialized design toolkit, MixITS-Kit.

Previous research has identified design patterns through a structured reflection on artifacts created during the iterative development of a MR industrial safety training system (Rauh et al., 2024). This approach aligns closely with our methodology, as both rely on post-project reflection on project documentation to extract design patterns relevant to the MixITS domain. However, their study does not address the unique potentials and challenges introduced by incorporating AI-driven instruction and feedback.

Valuable studies have devised guidelines at the intersection of MR and AI, which is closely related to MixITS-Kit haven't concentrated the use case of real environment task support. XAIR (Xu et al., 2023) is a design toolkit for explainable AI in augmented reality (AR). It was developed through a multidisciplinary literature review, surveys, workshops, and user studies. It offers guidelines for determining when, what, and how to explain AI outputs to AR users. While XAIR provides valuable insights for integrating explainability into MixITS systems, addressing factors like user cognitive states and environmental context, our toolkit takes a broader approach to MixITS system design challenges beyond explainability alone.

Following a review of 311 papers covering XR and AI, published between 2017 and 2021, Hirzle et al. (2023) identified research opportunities in the intersection of XR and AI. Our work builds on Hirzle et al.'s recommendations for future research by providing design tools to analyze human-AI interaction challenges in XR, specifically in the context of physical task guidance.

3 Methods

We aim to offer designers a structured toolkit for MixITS design that helps to tackle interaction problems between the user, system, and the real environment from multiple levels of abstraction. Prior research has leveraged widely available products in well-documented domains such as web search and recommendation systems to iteratively derive human-AI guidelines (Amershi et al., 2019). Unfortunately, this method is not directly applicable to the MixITS domain at the moment, due to the scarcity of documentation focusing on the design and the general availability of products in this space, as presented in Section 2.2. Another obstacle to the development of MixITS design tools is the

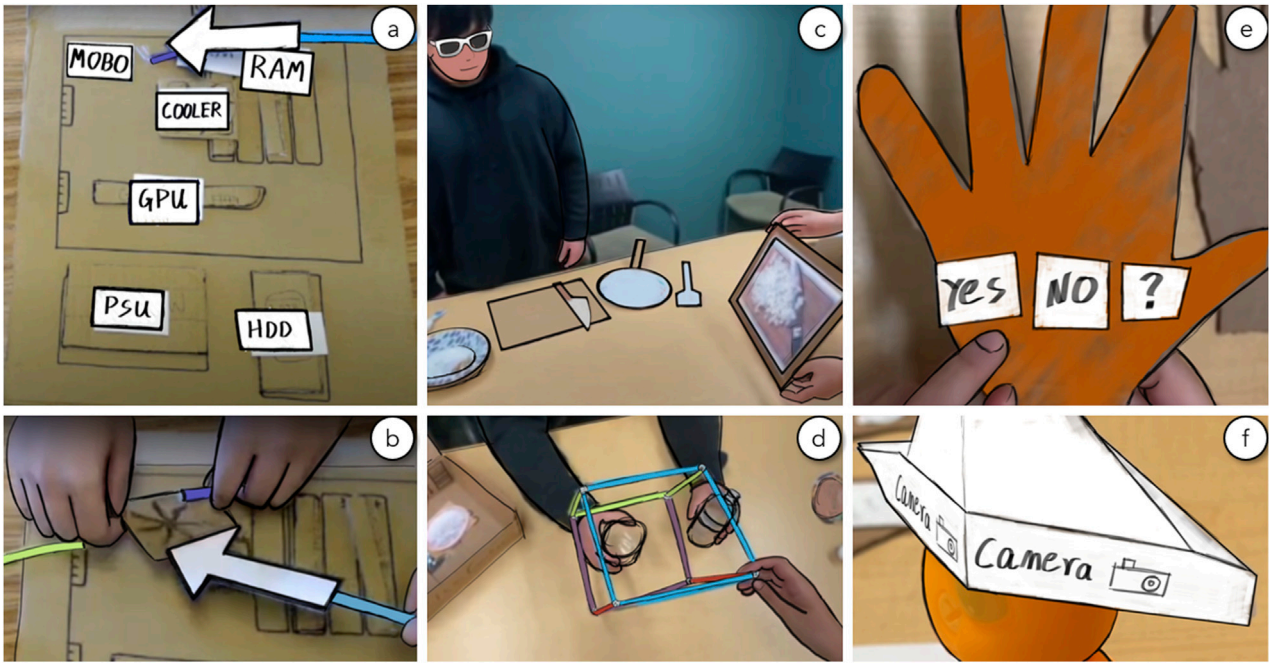

FIGURE 1
Low-fidelity prototyping allowed students to iterate on the design of MixITS systems with a small budget. Students created cardboard props representing real objects, such as a PC motherboard for AI-assisted repair **(a)** and kitchen utensils for AI-assisted cooking **(c)**. The MR interface prototypes used low-cost materials like paper labels to simulate virtual labels on a PC motherboard **(a)**. Students moved these props to prototype dynamic behaviors, such as a virtual arrow pointing to parts of a PC **(b)** and tracking ingredients with bounding boxes made of straws **(d)**, similar to low-fidelity paper prototyping of mobile or web applications. They also used inexpensive materials to represent customized hardware, including a haptic feedback glove **(e)** and a 360-degree camera headset **(f)**. To simulate large outdoor environments, such as a four-way stop for an AI-supported navigation app for blind users, students created miniature versions using dioramas, figurines, and toys with extra sensors such as cameras embedded in paper hats. This approach allowed them to enact full system functionality in a controlled, scaled-down setting that represented complex real-world scenarios **(f)**.

engineering complexity, long development cycles, and high costs associated with integrating AI into real-time MR applications.

To overcome the lack of availability of existing MixITS systems their development costs, we adopted a middle-ground approach using low-fidelity prototypes created by graduate students in a 10-week long course. We employed techniques like bodystorming (Burns et al., 1994; Oulasvirta et al., 2003), video prototyping (Leiva et al., 2020), and role-playing (Svanaes and Seland, 2004) to explore potential design challenges without full implementation. This method yielded valuable insights, leading to high-level design considerations and low-level solutions for identified problems presented as design patterns. Furthermore, the focus on problem definition and low-fi prototyping allowed us to propose design tools that are agnostic to the current technological limitations. We base our analysis methodology on prior research on design with AI (Lupetti and Murray-Rust, 2024) while addressing common challenges associated with using design workshops in research (Elsden et al., 2020).

We structured the design course to address common challenges associated with using design workshops in research (Elsden et al., 2020). By coupling in-class discussions and multiple prototype iterations, students engaged deeply with the material, resulting in higher-quality outputs than would be possible in shorter, broader-audience workshops. Students had dual motivations—enhancing their knowledge of human-AI interaction and MR while meeting academic requirements—which contributed to sustained engagement and meeting objectives over the 10 weeks. The structured assignment schedule with in-class presentations and feedback not only facilitated a continuous data collection process but also enabled systematic organization and allowed for deeper analysis of the evolving design prototypes, reflecting the iterative nature of the design process.

Class activities and assignments were designed with the pedagogical goal of educating future AI researchers about challenges in MixITS system design and exploring potential solutions. The curriculum aimed to help students understand the problem domain of human-AI interaction in MixITS systems, as well as the challenges and complexities involved in developing realtime, AI-supported interactive systems for task assistance. Our local IRB approved the use of the assignment deliverables as research data. Students provided informed consent for their assignments to be used for research purposes after the academic term ended and grades were submitted. One team of four members did not consent to their assignment data being used for research so we excluded their data from our analysis. Figure 1 shows the low-fi prototypes produced by students.

3.1 Participants

The course had 29 graduate students from computer science, engineering, and neuroscience at our university, divided into nine

teams—seven teams of three members and two teams of four members. Their diverse backgrounds produced teams with a balanced understanding of AI, MR, and human factors. All students were new to the MixITS application domain. We consider our student pool to be representative of early-career designers, engineers, and researchers new to the MixITS domain but with prior AI, XR, and HCI experience acquired through courses or research projects.

3.2 Apparatus

3.2.1 Syllabus and readings

The course syllabus included a variety of learning activities designed to provide a comprehensive understanding of the different aspects of designing MixITS systems. The activities included lectures on topics such as introduction to AI and HCI, design thinking, human-centered design, low fidelity prototyping techniques, AI interpretability and explainability, human-AI interaction, and ethics. In addition to lectures, the course included frequent in-class discussions on selected readings. Students wrote weekly reflections on assigned readings, responding to three specific questions about each reading. Students also commented on their classmate's reflections. The course included project-related checkpoints where students presented their progress in class and received feedback from their peers and the instruction team. Assignments were designed to align with the lectures and readings to encourage students to apply the material to their practical design challenges. A critical component of the class was interactive activities earlier in the quarter, including role-playing exercises with props, where students acted out physical tasks (e.g., making coffee), played the roles of an AI agent and an interface, to gain a hands-on understanding of the challenges and considerations involved in designing MixITS systems.

3.2.2 Role-played MixITS scenario

The in-class coffee-making demonstration and role-play scenario highlighted key challenges in AI-based MixITS systems, providing students with a hands-on understanding of the complexities involved in real-time AI instruction and feedback. Several limitations of MixITS systems became apparent, including action misinterpretation and inability to handle unexpected events not present in the training data. The system's failure to detect user errors, such as adding salt instead of sugar, was also highlighted. Timing and sequence issues in managing procedural steps, along with environmental variables affecting object and action recognition, further presented aspects to consider in developing MixITS systems. The demonstration revealed how user expertise levels and mental models can significantly influence their interactions with AI systems. Novices and experts may approach and respond to AI guidance differently based on their understanding and expectations of the system's capabilities. By confronting realistic AI limitations, students were better prepared to design realistic and robust MixITS systems in their projects.

3.2.3 Group project and assignments

The primary outcome of the course was a group project where students designed and created low-fidelity prototypes of their MixITS systems. The project was broken down into five bi-weekly assignments. Students used tools such as empathy maps, sketches, paper, and video prototypes of role-played scenarios for their design assignments, which they presented in class for feedback. They were expected to integrate the feedback into subsequent iterations of their prototype. Written assignments for each prototype stage involved design reflections, anticipated AI and user errors with potential solutions, interface design elements, user evaluation processes and outcomes, and more.

3.3 Procedure

Our human-AI interaction course focused on MixITS scenarios was 10-week long with lectures of 1 hour and 15 min twice a week. The syllabus included MR design, human-centered AI, and prototyping techniques suitable. Throughout the course, each team developed a low-fidelity MixITS system with an incremental and iterative process. At the end of the course they evaluated their prototypes with users. Students documented the entire process with weekly assignments, in-class presentations, written reports and recordings of user testing. This rich documentation was the empirical basis to develop the MixITS-Kit.

3.3.1 Assignment 1 (explore)

In Week 2, Assignment 1, students analyzed a freely chosen physical task using flowcharts and sketches that detailed the task steps, required tools, expected outcomes, and surrounding environment. Next, they created empathy maps to gain a deeper understanding of their potential audience, adopting a user-centered approach. Students then reflected on the challenges, benefits, and risks of deploying an AI system to support users in the physical task. Their reflections resulted in AI-based use cases to facilitate learning and improve users' skills along with suitable interaction modalities. This assignment laid down the foundations for the next iterations.

3.3.2 Assignment 2 (empathize)

The second assignment focused on two primary objectives: fostering empathy for potential users and deepening understanding of a MixITS task from both user and AI perspectives. This approach aimed to cultivate a more empathetic and comprehensive design process, considering the viewpoints of both the users and the AI system. Students participated in a role-play exercise to simulate the interactions between a user, an AI system, and the interface in a MixITS scenario (Section 3.2.2). This activity incorporated established HCI methods such as the think-aloud protocol (Ericsson and Simon, 1980) and bodystorming (Burns et al., 1994; Oulasvirta et al., 2003). Through this hands-on experience, students gained insights into the complexities of human-AI interaction in a MixITS system, including the reality of AI errors, and identified interface design challenges to address these errors. Understanding the interface design challenges focused on managing AI errors while balancing several key factors: preserving user agency, minimizing cognitive load, reducing user frustration, maintaining trust in the system, and preventing users from disregarding the AI's input altogether. This approach aimed to help students recognize the various elements that need to be considered while designing human-AI interfaces that effectively

support users without compromising the overall user experience or the system's utility. For their submission, students recorded a prototype walkthrough and provided a written reflection on their prototype. This combination of practical demonstration and analytical reflection allowed students to synthesize their learning and apply it to their design scenario.

3.3.3 Assignment 3 (prototype)

Assignment 3 built upon the insights gained from the previous two assignments, challenging students to prototype solutions for their MixITS scenario. Rather than designing entire systems, students narrowed their focus to critical aspects such as managing AI errors, addressing user perception and trust, and examining system design assumptions. Students created low-fidelity prototypes of MR interfaces, employing a Wizard of Oz technique (Kelley, 1983) to produce video prototypes. This approach allowed them to explore design concepts without committing prematurely to complex and time-consuming engineering, as pointed in prior work on MR prototyping (De Sá and Churchill, 2012). They detailed the physical tasks in their scenarios by visually mapping out step pre-conditions and task sequences, using the provided drip coffee-making scenario as a guideline. Through this assignment, students learned to appreciate the nuances of accommodating task variations and error detection in AI-assisted systems. They gained experience in balancing user needs with system capabilities, and understood the importance of considering execution uncertainties in their designs. This exercise reinforced the complexity of creating intuitive, effective, and trustworthy human-AI interactions in MixITS scenarios.

3.3.4 Assignment 4 (iterate)

Assignment 4 focused on the evaluation of the refined prototypes, which now incorporated a simulated AI backend. The assessment process involved instructors acting as users to test both the interface's effectiveness and its technical design. To challenge the system's robustness, instructors intentionally introduced errors during testing when instructions were unclear or exposed to safety risks (e.g., cooking burns), helping students uncover design challenges relevant to real-world applications of MixITS systems. Students were tasked with detailing their user study design including tasks and metrics, writing reflective essays on their findings, itemizing instructor feedback and brainstorming potential solutions to observed interface failures. This reflective exercise served to deepen their understanding of their designs and provided valuable insights for future iterations of the projects.

3.3.5 Final assignment (reflect)

The final design project culminated in a comprehensive report on the designed MixITS system. This report discussed the rationale behind their designs, explored user benefits for learning new skills or performing tasks, and offered a critical reflection on the system's strengths and weaknesses. Students also provided recommendations for future designers working on similar projects. This final assignment encapsulated the entire design journey, integrating insights from previous assignments and incorporating feedback from peers and instructors. The project served as a reflection of their growth and understanding in designing AI-assisted MixITS systems.

4 Data analysis

We analyzed all course data, including video prototypes and accompanying written documentation produced by each team for every assignment. We performed two types of analyses: 1) reflexive thematic analysis (Braun and Clarke, 2006) which resulted in six design considerations, and 2) design pattern elicitation (Winters and Mor, 2009; Retalis et al., 2006; Borchers, 2000) which resulted in the identification of 36 potential problems and example solutions.

4.1 Reflexive thematic analysis

We used the flexible qualitative approach of reflexive thematic analysis, as formalized by Braun and Clarke (2006). This method allows either inductive or deductive strategies (Braun and Clarke, 2024). We chose a deductive approach starting from goal-oriented codes taken from the metacognitive questions in the assignments. Braun and Clarke emphasize that reflexive thematic analysis can be effectively conducted by a single analyst, and that inter-rater reliability is not necessary for the method to be applied rigorously (Braun and Clarke, 2006; Braun and Clarke, 2024; Maxwell, 2010).

Our analysis followed the 6-stage procedure proposed by Braun and Clarke (2006), Braun and Clarke (2024). First, the analysis was conducted by one of the authors who did not have any classroom role or interaction with the student teams. This data analyst started by familiarizing themselves with the assignments. The initial set of codes was based on the metacognitive (Flavell, 1976) components in the assignments—design challenges, decisions, trade-offs, and recommendations for future designers. Following the initial coding, the analyst developed a set of themes. These were presented and discussed with the teaching team, to ensure they were coherent to the classroom experience and captured insights gained during the design course. The discussion resulted in a final set of themes that are named and reported in Section 5 as design considerations.

4.2 Design pattern elicitation

Drawing inspiration from previous work on design pattern elicitation (Winters and Mor, 2009; Retalis et al., 2006; Borchers, 2000), our method to elicit MixITS design patterns consisted of, (1) compiling case studies, (2) identifying common functionalities, (3) decomposing the functionalities into triplets of *context, problem, and solution,* and (4) refining the triplets by detailing, merging, splitting, and removing. The eight prototypes were used as case studies, represented by written reports and video recordings. We reviewed the reports and watched the videos for each prototype, to compile a coarse list of 63 functionalities identified across the eight prototypes. Each of the 63 functionalities was broken down into triplets of *context, problem, and solution.* Following that, the 63 triplets were labeled with one of the eight Interaction Gulfs presented in Section 5.2. The Gulf labels abstract the details of the context and for this reason the full context description was omitted from Table 1. The gulf labeling step involved identifying the actor (Human or AI) and the target (Human, AI, or the Environment) of

TABLE 1 The initial 36 MixITS Design Patterns in the Problem-Solution format tuple and grouped by Interaction Gulf.

Gulf	ID	Problem	Example Solutions
H-Ex-E *Human Execution on the Environment*	1	Environment Understanding	Provide additional information about a real environment element *E.g., descriptive labels registered to unfamiliar components of a PC.*
	2	Risk Awareness	Highlight potentially dangerous real environment elements before the user executes the step *E.g., MR arrows and audio warning about a hot pan nearby*
	3	Step Preconditions	Notify the user about an unfulfilled precondition of the next step *E.g., A spoken instruction to the user to tune their guitar before starting practice*
	4	Procedural Knowledge	Provide instructions with written text and speech *E.g., Speech synthesis and MR text/visuals instructing the user on how to operate a rice cooker*
	5	Spatial Knowledge	Visually demonstrate the trajectory of an action or the direction of the relationship between two elements *E.g., Animated MR arrow moving between sockets to guide PC cable re-connection*
	6	Body Movement	Visually demonstrate body movement or pose with an animated 3D body, video, or image *E.g., Animated MR hand to demonstrate how to play a lead passage on the piano.*
	7	High User Workload	Provide virtual functionalities that aid in the task execution *E.g., Timer to keep track of cooking time*
	8	Virtual Environment Clutter	Hide virtual overlay *E.g., A voice command to hide or dim MR labels on objects in a repair shop*
	9	Real Environment Clutter	Indicate real environment elements that can be removed from the workspace *E.g., Register MR labels to indicate a utensil is not needed for a recipe and can be stored*
	10	Lack of Focus	Highlight a single real environment element *E.g., Complement instructions by pointing an MR arrow to a PC component that should be connected*
H-Ev-E Human Evaluation of the Environment	11	Expected Task Result	Describe the expected outcome of an entire task *E.g., Image of the dish the user wants to prepare*
	12	Task Result	Provide a comprehensive report of the final product of the task, explaining inconsistencies and suggesting corrective actions *E.g., Superimposed MR model of a correctly assembled PC over the user's assembly to visually highlight discrepancies*
	13	Task Performance	Provide a comprehensive report of the user performance throughout the task, explaining mistakes and suggesting technique improvements *Line plot on an MR panel showing user tempo accuracy, highlighting mistakes with audio samples, and suggesting tailored training for those passages*
	14	Task Progress	Inform task completion progress *E.g., MR panel with a checklist of steps*
	15	Expected Step Results	Describe the expected outcome of a step of the task *E.g., Image of herbs diced or chiffonade as required by the recipe*
	16	Step Results	Inform the user of an inconsistency in the step outcome and recommend correction based on the impact *E.g., Interrupt (if allowed) the user during incorrect chord play and show an image of the correct music sheet.*
	17	Tool Operation	Demonstrate the correct operation of the physical tool and contextualize mistakes by overlaying content on the tool *E.g., Overlay MR numbered labels on the guitar neck to display the correct chord shape*
	18	Causal Chain	Review previous step records with the user to identify undetected circumstances that caused inconsistencies in outcomes *E.g., Engage in a conversation with the user to reason about pieces of evidence after analyzing a crime scene*
	19	User Frustration	Reassure the user, offer instructions to correct performance or outcome issues, and alternative courses of action *E.g., Synthesize reassuring speech and corrective measures in a calming tone after the user drops a bowl of ingredients*
	20	Spatial Knowledge	Visually demonstrate the trajectory of a previously executed step or the relationship between two elements involved in a previous step *E.g., Demonstrate the trajectory a suspect may have followed to enter a crime scene based on the evidence*

(Continued on following page)

TABLE 1 (*Continued*) The initial 36 MixITS Design Patterns in the Problem-Solution format tuple and grouped by Interaction Gulf.

Gulf	ID	Problem	Example Solutions
H-Ex-AI *Human Execution on AI*	21	Dismissal	Allow users to dismiss AI system functionalities easily *E.g., "Thanks, I can take it from here." – User*
	22	Activation	Allow users to request AI system functionalities easily *E.g., "CookGPT (activation keyword), how much olive oil should I use?" – User*
	23	Scoped Activation	Allow users to easily request AI system functionalities directly related to a virtual or real environment element *E.g., MR buttons, linked to music sheet sections, offering guidance upon tapping*
	24	Task Obstruction	Support interaction modalities that do not obstruct user interactions with the real environment *E.g., speech, eye gaze, adaptive UI, tangible inputs*
H-Ev-AI *Human Evaluation of AI*	25	Tracking in Progress	Inform the user about the parts of the environment being tracked and the expected end time during tracking *E.g., MR frame that moves along the environment as the system is scanning and reconstructing specific parts of the environment*
	26	Chain-of-Thought	Allow user-system dialogue by explaining the system's decision-making process and enabling user inquiries or counterarguments *E.g., Engage in a conversation with the user to explain the reasons why the system is suggesting an alternative theory to explain a crime*
	27	System Focus	Inform the user of the system's focus shift to a specific area within the physical environment *E.g., An MR icon around the user's hands, while they are mixing ingredients to communicate to the user the system, is aware of the action*
	28	Privacy	Request explicit user authorization for the acquisition of personally identifiable information *E.g., Clearly display the terms of use and obtain explicit user consent to collect data about the user's home*
AI-Ex-E *AI Execution on the Environment*	29	Act on the Environment	Coordinate with IoT devices to change the environment towards the user goal *E.g., The system requests a smart traffic light, after considering everyone's safety, to turn red so the user can cross the street.*
AI-Ev-E *AI Evaluation of the Environment*	30	Limited Spatial Awareness	Environment modeling from spatial data captured through multi-view RGB, depth-cameras, LiDAR, or millimeter wave radars *E.g., A 360-degree camera attached to the MR HMD.*
	31	Step Results	User feedback to improve the system's assessment of a specific task step outcome locally in the session *E.g., The AI misclassifies a chord as a mistake due to background noise, but after the user corrects the system, it recalibrates the microphone*
	32	Procedural Knowledge	Allow users to analyze outcomes of past instructions to correct the system's procedural knowledge base *E.g., The user informs the system about an incorrect ingredient in the recipe, prompting the system to update the recipe accordingly*
	33	Environment Model	Situated user feedback to correct mistakes in the system's environment model *E.g., hand gestures to annotate artifacts of misclassification or 3D reconstruction in the system's environment model*
	34	Task Progress	User feedback to update task progress in case of undetected step completion *E.g., The user informs the system via speech that they have already mixed the necessary ingredients, which was not previously detected by the system*
AI-Ex-H *AI Execution on Human*	35	Instructional Language	Rephrase instructions to facilitate understanding when users demonstrate doubt after receiving guidance *E.g., In a PC repair scenario, replace a technical term (CPU Cooler) with lay terminology (fan attached to the metallic grid)*
AI-Ev-H *AI Evaluation of Human*	36	Goal	Infer user goals based on explicit inquiries or implicit analysis of their actions in the real environment *E.g., Initially, the system assumes the user is playing song A, but upon hearing the first notes, it identifies the song as B and updates the instructions*

the interaction. Additionally, the actor's goal and the target's means to enable the user to accomplish their goals, in the execution cycle, and the actor's interpretation of the target's feedback, in the evaluation cycle, were labeled. Our experience, repeating this step for each of the 63 triplets, resulted in the MixITS Interaction Canvas, a design tool to support Interaction Gulf analysis, described in more detail in Section 5.4. After a deep reflection phase followed by discussion among the authors, we refined the labels and combined redundant triplets to generate a final set of 36 emerging design pattern triplets reported in Section 5.3.

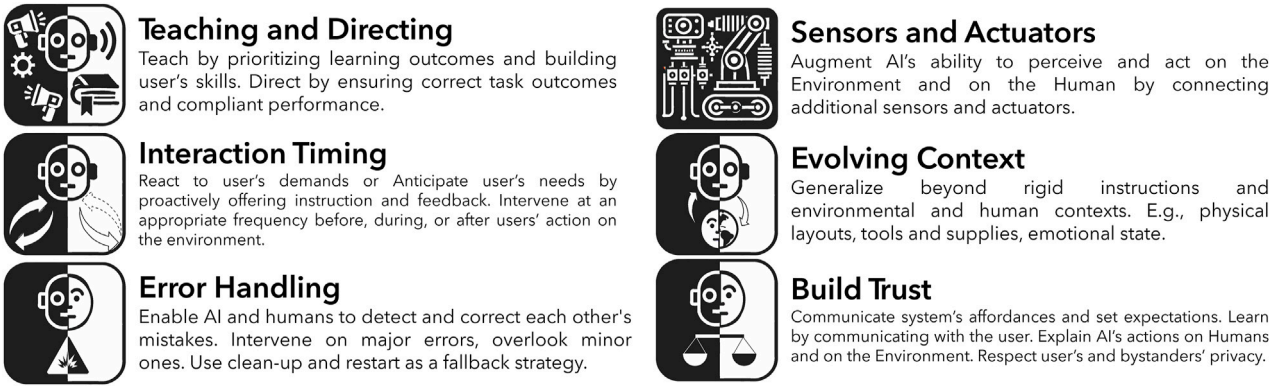

FIGURE 2
Six MixITS Design Considerations produced by a reflexive thematic analysis of data collected in the design course. Section 5 presents the considerations in detail.

5 MixITS-kit design toolkit

From the analysis of the extensive documentation of the interactive design process of MixITS prototypes during the course, we distilled six design considerations for the design of future MixITS systems. We encourage readers to consider the similarity of their project domain to MixITS (Section 3.3) to judge their "proximal similarity" (Campbell, 1986) before transferring our findings to their particular cases (Polit and Beck, 2010). The design considerations are detailed in the following subsections and summarized in Figure 2.

5.1 Themes and design considerations

The design considerations expand existing human-AI design guidelines by considering adaptive AI, multimodal reasoning, and context-aware interaction. These principles, while drawing from prior AI and MR guidelines, uniquely address real-time physical task support challenges. By emphasizing user state inference, environmental adaptation, and diverse interaction modalities, they bridge digital interfaces and physical tasks. The considerations recognize MixITS systems as operating at the intersection of AI, HCI, and physical performance, requiring nuanced application of cross-domain principles to enhance system efficacy and user experience.

5.1.1 Teaching and directing

We observed that design choices typically aligned with two main goals: teaching skills or directing tasks. Skill-focused MixITS aimed to empower users for independent future ability, implementing pedagogical interventions and assessing learning outcomes. Task-directing MixITS focused on either meeting outcome specifications or guiding users through precise steps. These two approaches differed in their user assumptions and environmental focus.

5.1.1.1 Teaching-focused

MixITS prioritized learning over task completion, sometimes interrupting activities for educational purposes. For example, PianoMix interrupted a user's session to correct technique errors, even when the technique did not adversely affect the musical output. ARCoustic modulated feedback and modeled user emotions to maintain motivation, reflecting research on engagement and learning outcomes (Harp and Mayer, 1997; Erez and Isen, 2002). These designs also allowed users to select instructional preferences and performance metrics self-assessment. Surprisingly, teaching MixITS often lacked safety features and assumed well-controlled environments. They also provided fewer opportunities for users to correct AI mistakes, presuming an "ideal" system state and superior AI knowledge.

5.1.1.2 Task-directing

MixITS assumed higher user expertise in real-world settings. These designs allowed users to provide feedback on AI mispredictions (RealityFix and ChefMix), and were more considerate of AI interruption frequency and task disruption. Some systems prioritized task continuity over immediate error correction. User agency was preserved by allowing reasonable improvisation within task steps (CrossReality and ARCoustic). Intervention decisions resembled those proposed by Horvitz principle (Horvitz, 1999). Safety was a key focus, with designs signaling potential hazards (Cookbot and ChefMix). CrossReality even included the ability to request external medical assistance.

When designing MixITS systems, it is crucial to distinguish between teaching and directing goals, as this choice significantly impacts design decisions. Teaching MixITS should consider prioritizing learning outcomes, ensuring safety even in controlled environments, and incorporating pedagogical interventions (Choi et al., 2014; Laurillard, 2013) derived from research on Intelligent Tutoring Systems (Brown and Burton, 1978; Anderson et al., 1985). In contrast, directing MixITS should consider balancing task continuity with error correction, preserving user agency within task parameters, allowing users to correct AI mistakes, focusing

on real-world safety, and document processes for future auditing and improvement, especially in industrial contexts. Both types of systems need to consider AI interruption frequency, user expertise levels, and feedback mechanisms. While incidental learning may occur in task-directing MixITS (van Asselen et al., 2006; Tresselt and Mayzner, 1960), clearly defining the primary goal as either teaching or directing would lead to more optimized and effective designs tailored to their specific purposes.

5.1.2 Interaction timing

MixITS systems were designed with either proactive or reactive interaction styles. Proactive systems offered unsolicited guidance, while reactive ones responded to user requests. This choice was influenced by user expertise, some metric of criticality (e.g., user safety and impact on the task outcome), and perceived user workload. For instance, a system for blind users proactively intervened in dangerous situations (CrossReality), whereas a piano-learning system reactively provided feedback when requested (PianoMix). Students recognized the challenge of balancing proactive interaction frequency with user agency. To address this issue, designs like MusIT, a guitar-teaching prototype, allowed users to adjust the intervention frequency according to their preferences. Students designed interventions for three key phases: before, during, and after tasks. Before-task interventions typically included feature tutorials (ChefMix) and procedure summaries (Cookbot and ARCoustic). During-task interventions ranged from minimal guidance (PianoMix and ARCoustic) to critical mistakes or safety alerts (*Cookbot*). Post-task interactions offered performance reports and outcome assessments. To address interaction frequency concerns, some designs, like MusIT, accumulated recommendations for periodic or end-of-session reports, demonstrating how temporal alignment can optimize user experience.

Designers should consider balancing proactive and reactive interactions and their frequency based on user expertise, task criticality, and workload. Simultaneously, they need to be mindful of how interaction frequency impacts user workload, fatigue, and error rates. A flexible system would allow users to adjust AI intervention frequency and tailor interactions to specific task phases, which can help personalize the experience and retain user agency.

Optimizing interaction frequency can avoid disrupting task flow while maintaining system usefulness. System actions, if aligned with task execution, would minimize disturbance, and implementing feedforward techniques for "before-task" interactions would help users anticipate consequences. Interactions that are too infrequent might lower the system's usefulness with suboptimal results and inefficiencies. This resonates with the guideline to "time services based on context" by Amershi et al. (2019) for non-MixITS systems. Existing guidelines on mixed-initiative systems (St. Amant and Cohen, 1997; Allen et al., 1999), utility concepts (Horvitz, 1999), and contextual timing of services offer valuable insights that complement our design considerations. The implications of unsolicited assistance for learning have been explored in the intelligent tutoring systems domain as the "assistance dilemma" (Koedinger and Aleven, 2007). Frameworks such as XAIR (Xu et al., 2023) can provide specific guidance on timing AI explainability effectively.

5.1.3 Error handling

Students identified errors in user-environment interactions via a simulated walkthrough with a task flowchart analysis that highlighted steps, outcomes, and required resources. Cognitive walkthroughs and prototype testing revealed potential errors in user-AI interactions and AI's perception of the environment. This process uncovered potential errors in MR interface interactions, AI predictions, and environmental changes. Four themes emerged from students' error-handling strategies.

5.1.3.1 Prevention

Prototypes demonstrated strategies involving users in decision-making. MusIT incorporated a feedback loop where the AI prompted the user for confirmation of critical information before performing tasks like optical character recognition (OCR) on user-shared documents. ChefMix had two error prevention features: one allowing users to notify the system about unfulfilled step preconditions before executing an AI-instructed step, and another where the AI preemptively detected unfulfilled preconditions overlooked by users. These approaches showcase collaborative user-AI error mitigation methods.

5.1.3.2 True positives

Teams implemented various strategies for handling AI and user errors. When AI detected its own mistakes, RealityFix acknowledged and apologized, while Cookbot implemented a collaborative problem-solving approach initiated by the user describing the problem. For user errors detected by AI, PianoMix and ARCoustic could interrupt tasks and provide immediate corrections, ensuring proper practice. ChefMix went further by incorporating emotional state management before guiding users through corrections, aiming to reduce frustration and enhance the learning experience. These approaches indicate the goal to increase transparency, build user trust, and improve error handling in AI-user interactions.

5.1.3.3 False positives

We identified strategies for handling AI false positives in error detection. PianoMix enabled users to dismiss incorrectly identified errors, maintaining user control over task disruptions. ARCoustic implemented an interactive approach where users could counter-argue AI-detected mistakes and add an explanation to correct the AI's understanding, offering a way to override the system's false positives and provide data for future refinement of the algorithm. Notably, no features were identified where AI could judge user-detected errors as false positives, suggesting a prioritization of user agency in error handling. These approaches indicate a balance between AI assistance and user control in error management.

5.1.3.4 Intractable errors

Teams devised various strategies to handle errors that went undetected by both users and AI during the task. Cookbot offered a user report feature for post-task issue reporting to developers, which can assist future system updates. RealityScan had an AI feature supporting users in deductive thinking, enhancing error log quality through collaborative error detection. Fail-safe mechanisms were included on Cookbot and ChefMix, allowing users to manually shut down the system, clean the work area, and start again. CrossReality

focused on user safety, incorporating shortcuts and safeguards to trigger external support like medical help when necessary.

5.1.3.5 Design considerations

Designers should consider prioritizing error prevention and comprehensive error handling. This could be done by incorporating proactive strategies, such as feedback loops for user confirmation of critical information and preemptive detection of unfulfilled preconditions for a specific task. Feedforward techniques that allow users to anticipate the consequences of their actions have been shown to significantly decrease errors by enhancing user understanding of the system's behavior and potential outcomes (Coppers et al., 2019; Muresan et al., 2023; Artizzu et al., 2024). Error handling mechanisms need to address both AI and user mistakes, including AI acknowledgment of its errors, user-initiated problem-solving, AI interruption for immediate user error correction, and related impact on learning and flow. Emotional regulation features can enhance the learning experience by reducing user frustration during corrections. The design should consider user agency in dismissing false positives and the ability to provide feedback to the AI without making it burdensome. Post-task error reporting could help address undetected issues while helping improve the overall system. Depending on the context, fail-safe mechanisms like manual system resets and safety triggers for external support can be integrated. Overall, there is need for designs to balance AI assistance with user control, prioritize transparency to build trust, and ensure adaptability to various error scenarios in physical task contexts. With a human-in-the-loop, designers can enable users to guide and correct AI behavior in realtime, thus enhancing the system's adaptability and reliability over time. Designers can additionally leverage prior research on human errors that provides useful taxonomies and approaches to handle skill-based, knowledge-based, and perceptual errors (Shappell and Wiegmann, 2000; Norman, 1981).

5.1.4 Sensors and actuators

Some MixITS prototypes showcased enhanced sensing and actuating capabilities beyond standard MR headsets. ChefMix installed fixed cameras in the environment, providing a third-person view to reduce occlusions and improve physical environment modeling. CrossReality, focusing on assisting blind users in street crossing at a stop sign, developed a system that coordinated with smart city devices like traffic lights and vehicles. Their prototype included wearable devices using haptic stimuli for guidance and an augmented cane with buttons for user input, enabling effective interactions in noisy environments without relying on audio feedback. These enhancements show the potential for MixITS systems to integrate with broader infrastructures and adapt to diverse user needs and environmental challenges.

Future MixITS systems should consider integrating IoT and robotics to enhance environmental interaction capabilities. Augmented tools with embedded sensors could provide precise task guidance, while leveraging existing 3D mapping data could enhance spatial awareness. A wearable ecosystem could provide physiological data for personalized guidance, and networked expertise could offer real-time access to specialized knowledge and potentially reduce AI errors. Bodily control technologies like Electrical Muscle Stimulation (EMS) (Nith et al., 2024; Shahu et al., 2023) and Galvanic Vestibular Stimulation (GVS) (Sra et al., 2017; Byrne et al., 2016) could add to user path guidance, potentially reducing workload and enhancing performance. Designers need to address challenges such as sensor obstruction, particularly as MR headsets block facial expressions, necessitating alternative methods for emotional state inference. Leveraging existing technologies and infrastructure can help overcome current limitations in AI and MR technologies, enabling more sophisticated user-environment interactions in task guidance.

5.1.5 Evolving context

One of the major benefits of MixITS systems over other types of guidance such as video tutorials, beyond realtime feedback, is their potential for adaptability. MixITS systems are expected to generalize beyond rigid instructions and seamlessly adapt to the user's context and ability. ChefMix and ARCoustic used human reactions as a shortcut to infer contextual changes, bypassing pure machine perception. They detected important changes through shifts in facial expressions and gaze patterns, prompting proactive system responses. The prototypes also adapted their interaction modes to environmental changes. For instance, ARCoustic typically provided feedback after listening to a user's guitar performance, but in noisy settings, it temporarily switched to demonstrating song passages and allowing for user self-assessment.

AI can allow systems to learn and refine based on user experience and feedback (Ouyang et al., 2022). When designing MixITS systems, designers should consider: implementing adaptive learning based on user feedback; utilizing multimodal AI to process various data types for comprehensive reasoning; inferring user workload and interaction capacity through multimodal analysis (Liu et al., 2024); detecting changes in manipulated objects for contextual adaptation; and leveraging diverse hand grasps to enable a range of interaction possibilities ranging from single-finger gestures to graphic grasp-centered interfaces (Sharma et al., 2021; Aponte et al., 2024). These considerations allow for more responsive, context-aware systems that can adjust their guidance methods based on user needs and environmental factors.

5.1.6 Building trust

In the context of physical task guidance, research indicates the importance of trust for effective human-AI interaction (Haesler et al., 2018; Eyck et al., 2006). Students addressed this by incorporating trust-building features into their prototypes. Cookbot, ARCoustic, and MusIT allowed users to report AI mistakes, enabling system adaptation and fostering a sense of user influence. ARCoustic and MusIT explained errors and recommendations to the users, while Cookbot and ChefMix used MR visuals to show tracked environmental elements, enhancing transparency. These approaches, aligned with frameworks like XAIR for explainable AI in AR (Xu et al., 2023), aimed to establish two-way communication between users and the system, to foster long-term trust development. Explainability is a key factor in trust-building, that aligns with Amershi et al.'s guideline to clarify system actions (Amershi et al., 2019). Cookbot and ARCoustic visualized task progress and allowed users to fast-forward or return to a specific step of the task procedure, enhancing user control. Addressing

privacy concerns, students recognized the sensitive nature of multimodal data collection in MixITS systems. PianoMix implemented informed consent procedures, while other projects proposed visual indicators for active data collection and measures to protect bystander privacy, demonstrating awareness of ethical considerations in AI-driven task guidance.

While not unique to MixITS systems, designers should focus on building trust through transparent decision-making, error reporting, and system adaptation. Explainable AI using visual cues to demonstrate awareness and reasoning, should be considered. Enhancing user control with task navigation and progress visualization can help with user agency and associated trust. Prioritizing privacy through clear consent procedures and data collection indicators can help users feel comfortable using a MixITS system. Clearly communicating system capabilities and limitations can help set accurate user expectations, particularly as questions about artificial general intelligence arise (Terry et al., 2023). This consideration resonates with the guideline to "Make clear what the system can do" from (Amershi et al., 2019). Designing adaptive systems that adjust guidance methods based on user feedback and environmental factors can help ground the systems into the user's immediate context. These considerations can create trustworthy, transparent, and effective MixITS systems that balance advanced capabilities with user-centric design.

5.2 Gulfs of execution and evaluation in intelligent mixed reality

Donald Norman's Gulfs of Evaluation and Execution (Hutchins et al., 1985; Norman, 1986; Norman, 2013) have been instrumental for HCI researchers and practitioners in understanding the challenges users encounter when interacting with systems (Vermeulen et al., 2013; Hornbæk and Oulasvirta, 2017; Muresan et al., 2023). The gulf of execution refers to the gap between a user's intended action and the options a system provides to perform that action. A small gulf indicates that the system's controls align well with the user's intentions, making it easy to use. A large gulf of execution means the system's interface poorly matches the user's goals, making it difficult to accomplish tasks.

The Gulf of Evaluation is the gap between how a system presents its state and how a user interprets it. A small gulf of evaluation means users easily understand the system's feedback and current state. A large gulf of evaluation indicates users struggle to accurately interpret the system's state, leading to confusion or errors.

We apply Norman's gulfs to identify and analyze interactions in a MixITS system, which includes the new element of the physical environment and changes the system from the traditional deterministic system to an AI-based probabilistic one. In a MixITS system, users need to convey their intentions to both the AI (system) and the physical environment simultaneously. This dual interaction can create two distinct gulfs: (1) Gulf of Human Execution on AI (H-Ex-AI) occurs when users struggle to communicate their intentions to the AI system, and (2) Gulf of Human Execution on Environment (H-Ex-E) which arises when users face challenges in physically executing their intentions in the real world. For example, in an AI-assisted rock climbing system, a climber might intend to reach a misidentified hold on a path suggested by the AI. They need to indicate this mistake to the AI (H-Ex-AI) to allow it to update its suggested path, perhaps through gaze or gesture, while also physically moving their body to grasp a different hold (H-Ex-E) that is more within reach.

After taking action, users must interpret feedback from both the system and the physical environment. This dual interpretation can lead to two distinct gulfs: (1) Gulf of Human Evaluation of AI (H-Ev-AI) which occurs when users struggle to understand or interpret the feedback provided by the AI system; and (2) Gulf of Human Evaluation of Environment (H-Ev-E) which arises when users face challenges in perceiving or interpreting the results of their actions in the physical world. Continuing the rock climbing example, after reaching for a hold, the climber needs to process the AI's feedback about their technique or next move (H-Ev-AI), perhaps displayed through MR visuals or audio feedback. Simultaneously, they must evaluate their physical position and stability on the wall (H-Ev-E).

We can consider MixITS systems as mixed-initiative systems (Horvitz, 1999) where the AI backend can behave as an agent and initiate interactions with the user and the physical environment. This mixed-initiative perspective leads to four additional Gulfs: Gulf of AI Execution on Human (AI-Ex-H), Gulf of AI Execution on the Environment (AI-Ex-En), Gulf of AI Evaluation of Human (AI-Ev-H), and Gulf of AI Evaluation of the Environment (AI-Ev-E). Figure 3 shows all the eight Gulfs of a MixITS application.

We treat the physical environment as a passive entity affected by user and AI actions, rather than as an active participant. This simplification keeps our design toolkit focused and practical, avoiding the complexities of fully modeling environmental factors.

5.3 Design patterns

Novice MixITS designers may struggle to apply general AI and MR guidelines to specific projects. Without a shared vocabulary between designers and AI developers, there is a risk of misinterpretations, redundant problem-solving, and inconsistent designs, impacting user memorability and experience Winters and Mor (2009). We propose design patterns to bridge this gap, offering concrete solutions as examples that can help mitigate the "cold-start" problem Winters and Mor (2009) and facilitate designers and researchers to build upon and expand our set.

We chose not to report pattern frequency to avoid misleading readers to think some patterns are more relevant than others (Maxwell, 2010). Our focus is on establishing a taxonomy of key challenges in MixITS design and suggesting examples of AI and MR solutions. We encourage the reader to determine which patterns are transferable to their needs (Campbell, 1986; Polit and Beck, 2010). The 36 patterns, grouped into eight interaction gulfs presented in Table 1, offer a foundation for future refinement as the MixITS community grows.

5.4 Interaction canvas

In the design pattern elicitation process (Section 4.2), we classified 63 interaction problems into one of eight interaction gulfs (Section 5.2). This exercise led to the creation of the Interaction Canvas (Figure 4), a visual tool to streamline the analysis of Gulfs of Execution and Evaluation in MixITS

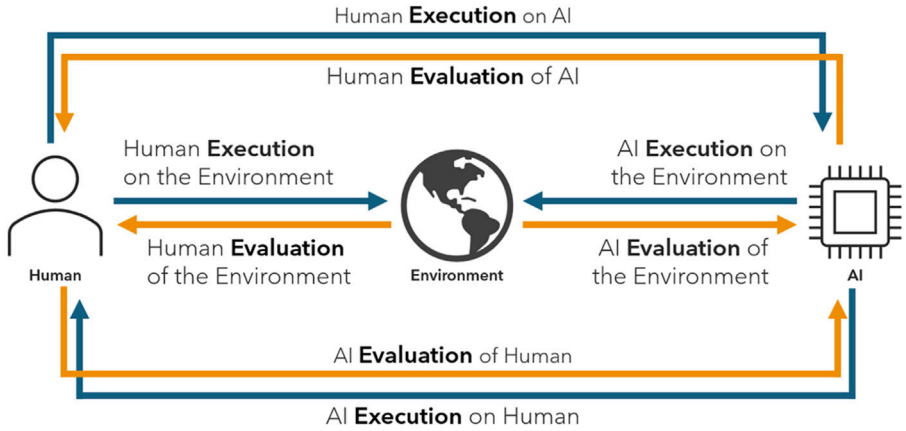

FIGURE 3
When applying Don Norman's Gulfs of Execution and Evaluation (Norman, 1986; Norman, 2013) to MixITS systems, we identify eight Gulfs that occur during interactions between the Human user, AI-MR system, and the Real Environment.

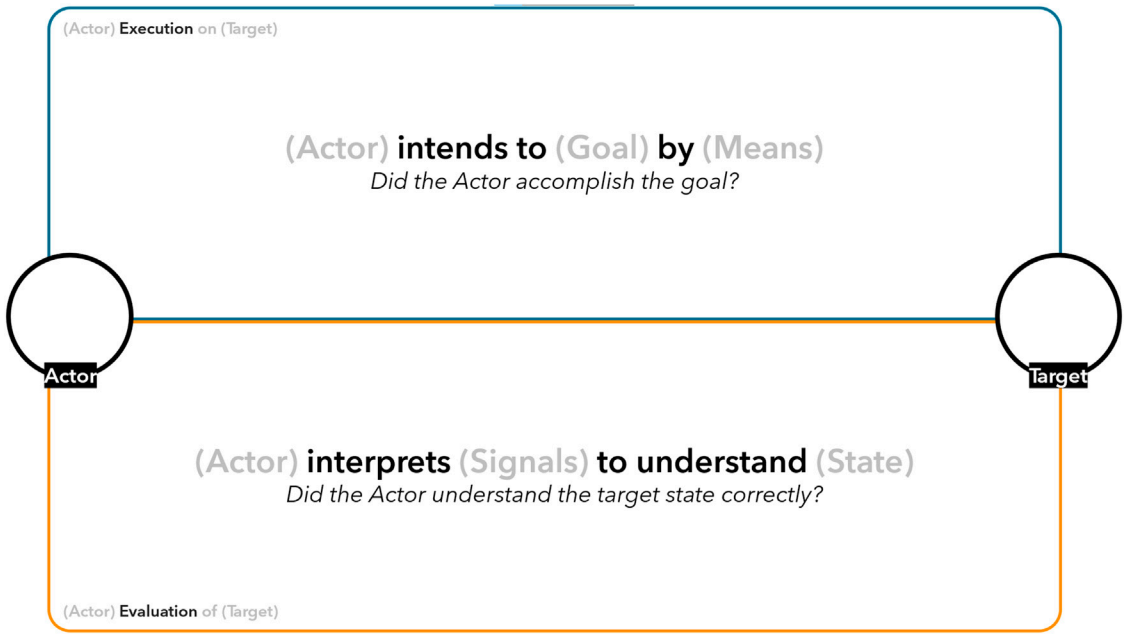

FIGURE 4
The MixITS Interaction Canvas can help designers identify interaction problems leading to the Gulfs of Execution and Evaluation between the AI system, Human users, and the Environment. Designers can use this visual tool by filling out the blanks, in gray text for both the questions and the gulfs.

systems. We present this Canvas as a tool to help designers visualize their thought processes and communicate effectively when analyzing interaction gulfs in MixITS systems. To use the Canvas, designers fill in the blank spaces at the edges of the canvas and answer: (1) who is the actor initiating the interaction, the AI-MR system or the human? (2) what is the target of the interaction, the AI-MR system, the human or the environment? Once the actor and target of the interaction are defined, the designer should focus on the execution cycle at the upper half of the Canvas and reflect on: (3) what is the actor's goal in the target? (4) what are the means provided to the actor by the target? (5) did the actor accomplish the goal? Next, designers can focus on the bottom half of the Interaction Canvas and answer: (6) what are the feedback signals emitted by the target? (7) How did the actor interpret those signals? (8) Did the actor understand the target state correctly?

6 Evaluation

To evaluate MixITS-Kit, which comprises the six design considerations, 36 design patterns, and the interaction canvas, we

conducted an asynchronous take-home study (Ledo et al., 2018) with eight participants. They were required to complete three tasks using the different components of MixITS-Kit to improve a fictional MixITS app for novice indoor rock climbers (ClimbAR). We gathered data from their work, including artifacts, feedback, and survey responses. The study was approved by our local IRB and participants provided informed consent before proceeding with the study tasks. Participants were compensated for their time.

6.1 Participants

We recruited eight participants (4F, 3M, 1NB; age range 18–33), all students from different departments of our home university who had prior experience with MR and AI. None of the participants were part of the human-AI interaction design course used for our initial data collection. We consider this participants are representative the target audience of MixITS-Kit, early career engineers and designers who have some technical experinece but are new to the MixITS domain.

One was a master's student, two were PhD students, and five were undergraduate students. Participants self-reported their expertise levels in designing AI, MR, and MixITS systems using a Likert scale (1 = "Never designed such a system," 5 = "Expert"). The mean self-reported expertise levels were as follows: 2.4 in AI (SD = 1.1), 2.7 in MR (SD = 1.2), and 2.0 in MixITS (SD = 1.0). We asked participants about their awareness of existing design toolkits for AI, MR, and MixITS. For AI, the most frequently mentioned toolkits were PyTorch (mentioned four times), TensorFlow, scikit-learn, and the Gemini API (each mentioned twice). For MR, the most commonly mentioned toolkits were MRTK (four times) and Unity AR Foundation (5 times). When asked about design toolkits for MixITS, participants either did not mention any (five times) or mentioned technical MR toolkits (three times).

6.2 Apparatus

The study was conducted remotely and asynchronously. Participants completed the tasks using a popular online slide authoring tool. We shared a digital version of our MixITS-Kit with participants using the same online platform. The Interaction Canvas was shared as a slide embedded in the main study slide deck containing Figure 4, the design patterns were shared as a PDF version of Table 1, and the design considerations were shared as slides with visuals and summaries from Section 5. We collected feedback and surveys using forms in the same online platform.

6.3 Procedure

Our procedure combines a walkthrough demonstration with a take-home study, similar to previous toolkit evaluations (Ledo et al., 2018). We instructed participants to work directly on the slides so all the changes were tracked. Each of the eight participants were randomly assigned to a unique condition starting at one of the eight MixITS gulfs and later encountering two other conditions varying gulf direction and actors least one change in the gulf direction and actor. This balancing strategy ensured every participant encountered every actor and every gulf direction.

Before participants proceeded with the actual study tasks, they were presented with introductory slides. These included one slide about Don Norman's principles of design (Norman, 2013), one slide with the definition of MixITS, and one slide describing the tools in our MixITS-Kit. Following this introduction, additional slides introduced the fictional setting of the study. This warm-up phase concluded with thirteen slides providing a walkthrough demonstration (Ledo et al., 2018) of how to accomplish a task similar to the upcoming task using MixITS-Kit. Each participant received a unique warm-up example different from the subsequent experimental tasks.

In the first task (Task 1), participants received a fictional error reported by the user. We created errors contextualized in the experimental scenario and based on one of the eight MixITS interaction gulfs according to the condition (Table 1). Participants described the user's issue by filling out the canvas (Figure 4), selected and justified a related interaction problem from Table 1, and proposed a solution using the associated design pattern. Participants detailed the solution by including text descriptions and a sketch image. This task assessed whether participants could effectively navigate design pattern catalog to identify a similar problem, use the canvas to analyze that problem, and apply a suitable design pattern to solve the reported issue. Task 1 is a realistic problem-solving scenario aided by our design toolkit, which is a relevant real-world UX design task suitable for our evaluation goals (Ledo et al., 2018; Alexander et al., 1977).

In the next task (Task 2), participants reviewed Task 1 solutions from other participants, ensuring they faced a scenario not previously encountered in either Task 1 or the warm-up phase. Given only the solution text description and sketch image, they had to identify which design pattern from Table 1 their peer applied, which then would help identify the associated gulf and problem. This exercise aimed to assess whether participants could recognize patterns in their peer's work, indicating the potential of MixITS-Kit as a shared language for MixITS designers and developers. Task 2 is a realistic pattern-matching exercise, requiring participants to interpret an existing solution using the toolkit's vocabulary—a practical UX design task aligned with our evaluation goals (Ledo et al., 2018; Alexander et al., 1977).

In the final task (Task 3), participants revised their Task 1 solution using a design consideration (Section 5) most relevant to their original scenario. All considerations were used at least once, with "Interaction Timing" and "Error Handling" used twice due to their complexity. This task aimed to explore how the design considerations can influence and creatively expand MixITS solutions. Participants completed a survey after this task. Task 3 involves creatively iterating on an existing solution using our toolkit, fostering essential UX design skills and supporting our evaluation goals (Ledo et al., 2018; Alexander et al., 1977).

6.4 Data collection

In the productive tasks (Tasks 1 and 3), participants were asked to rate their level of agreement with various statements using a

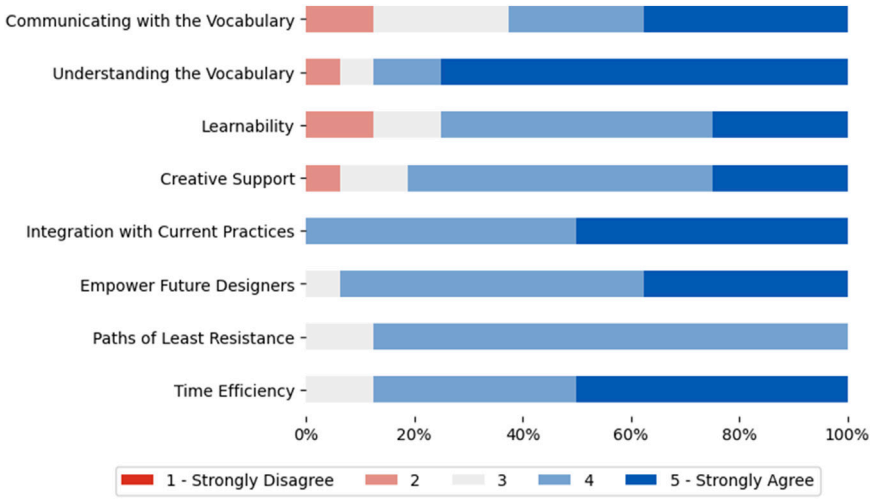

FIGURE 5
Participant agreement levels (5-point Likert scale) with statements about the Design Toolkit.

5-point Likert scale. The statements focused on Learnability (*"It was easy to learn how to use the toolkit."*), the toolkit's effectiveness as a shared vocabulary (*"I can easily communicate using the vocabulary of the toolkit."*, *"I can easily understand the vocabulary of the toolkit."*), and the research goals of the toolkit as proposed by Ledo et al. (2018) (*"It would take me longer to solve the task without the toolkit," "The toolkit helped me to identify paths of least resistance in the design process," "The toolkit will help to empower future designers," "The toolkit integrates well with current practices in design," and "The toolkit helped me to create a novel design."*). We also asked participants whether they found the toolkit too abstract and high-level, too low-level, or at an appropriate level of abstraction for productive tasks 1 and 3. We measured task completion time by summing the time intervals participants spent on tasks 1 and 3, as recorded in the slide change history. Task 2 evaluated participants' ability to correctly identify design patterns used by other participants. Therefore, we recorded only the names of recognized patterns and whether they matched the originally intended ones.

6.5 Data analysis

We analyzed the medians and median absolute deviations of the questionnaire responses and task completion times. Additionally, we examined the ratios of responses from the questionnaires, toolkit abstraction level feedback, and the correct recognition ratio from Task 2. One of the authors analyzed the solution descriptions and visuals from tasks 1 and 3 using the same pattern mining method employed in Section 5.3.

6.6 Results

The median task completion times in minutes were: Task 1 (M = 27.0, MAD = 9.0), Task 2 (M = 1.0, MAD = 0.5), and Task 3 (M = 8.5, MAD = 3.5). Participants generally responded positively to MixITS-Kit, as shown in Figure 5. Participants recognized the potential of MixITS-Kit to serve as a shared vocabulary among designers. However, they found it harder to communicate using the vocabulary (M = 4, MAD = 1) than to understand it (M = 5.0, MAD = 0.0). Participants considered it easy to learn MixITS-Kit (M = 4.0, MAD = 0.5). Regarding the level of agreement with sentences related to (Ledo et al., 2018) toolkit research goals, participants rated creative support (M = 4.0, MAD = 0.0), integration with current design practices (M = 4.5, MAD = 0.5), empowering designers (M = 4.0, MAD = 0.0), supporting paths of least resistance (M = 4.0, MAD = 0.0), time-efficiency (M = 4.5, MAD = 0.5). Regarding the level of abstraction, one of the participants (P9) considered the toolkit too abstract for the task (12.5%), in contrast to the other seven, who considered the components to be at appropriate levels of abstraction (87.5%).

In Task 1, two of the eight participants applied design patterns from gulfs different than anticipated and provided justifications that did not align with the defined gulf concepts, indicating a mistake. In Task 2, all participants correctly identified the actor and target in their peers' solutions. Four participants successfully recognized the correct design pattern, while five identified the correct gulf. One participant misidentified the pattern, and three incorrectly identified the gulf. Overall, half of the participants accurately identified the exact design pattern used in their peer's solution.

Participants identified the most challenging aspects of learning the canvas and design patterns for Task 1, including *"It was harder to identify the gulf."* - P9 and *"Understand the mapping between the gulf concepts and the terms in the toolkit"* - P10. When asked for open-ended feedback on the canvas and design patterns in Task 1, participants said, *"I thought the toolkit was very useful to standardize problems in AI-MR apps, and the design patterns seem pretty comprehensive to me."* - P1, *"adding an 'Other' section would allow developers to potentially communicate about rarer cases"* - P4, *"An interface that integrates both parts would make it easier to cross-reference."* - P6, *"I used the problem list to identify a*

problem that matched that one in the case, and then that helped me understand better how to fill each gulf" - P10.

When asked for open-ended feedback on the design consideration in Task 3, participants said, *"I think Task 2 and 3 were a lot more straightforward than the first task! After gaining familiarity with the design recommendation, I also found it easier to proceed."* - P8, *"Maybe it's just me being not creative enough, but it's still a little bit difficult to come up with a very novel design using the guidance in my scenario"* - P7 ("Sensors and Actuators" and Human Execution on the Environment), *"I used the design recommendation to find a pattern that aligned with the goal stated on it."* - P6, *"I thought the design recommendations were great guidelines to help find solutions to the proposed problems. It was especially interesting that they addressed both the AI and XR aspects of design, since most toolkits I've seen/used only work with one of those aspects."* - P1.

7 Discussion

MixITS-Kit consists of three interconnected components: an Interaction Canvas for analyzing execution and evaluation gulfs, a set of six user-centered design considerations, and a set of 36 low-level design patterns with example solutions. This structure can support designers throughout the development process, from conceptual analysis to practical implementation. Here we discuss our results, the synergy between the toolkit components, and their overall implications for the landscape of HAI.

7.1 Positive impacts of MixITS-Kit

One of the main findings from our evaluation was the high level of performance demonstrated by participants. Six out of eight participants successfully completed Task 1. In Task 2, all participants correctly identified the actors and targets, while half accurately recognized the design pattern used by their peers. P8 reported increased self-efficacy in Tasks 2 and 3 compared to Task 1, attributing this improvement to their growing familiarity with the process. Participants achieved these results with only a brief demonstration and minimal instruction during the 1-h evaluation. Given the inherent complexity of the MixITS domain and the challenges faced by participants in the formative class, we consider these results promising. This indicates that MixITS-Kit effectively distills prior design experiences, guiding novice designers through the complexities of MixITS.

Our evaluation results show general participant agreement that MixITS-Kit achieved the toolkit research goals proposed by Ledo et al. (2018). We attribute these results to the combined use of the toolkit components that helped participants to tackle MixITS design problems from different levels of abstraction. The design patterns with example solutions provided an accessible starting point, helping participants overcome the "cold-start" (Winters and Mor, 2009) problem and begin iterating on their own solutions earlier. P6 effectively used the problem column of the catalog to find a design pattern that matched their case in Task 1. P1 highlighted the value of having standardized problems in the AI-MR domain. These observations emphasize the value of our cataloged problems, perhaps more so than the solutions, as solutions may significantly change with technology, but the core problems remain consistently relevant and future-proof.

As illustrated in Figure 6, P6 utilized descriptive AR labels (H-Ex-E-1) with instructional guidance on crashpad importance, aligning with "Teaching and Directing." After considering "Build Trust," P10 incorporated an AI-generated explanation for the goal inferred by AI (AI-Ev-H-36). These are examples that the high-level considerations can inspire changes in previously proposed solutions. We speculate that takin our considerations in the early stages of MixITS development can strengthen the design and avoid drastic changes later on, possibly reducing rework.

Participants 1 and 6 observed strong synergy among the toolkit's components. By approaching the design patterns with a specific consideration in mind, they successfully identified a solution. For the problem in Task 1, P6 utilized the design pattern catalog as a reference to complete the canvas. These examples illustrate how designers can benefit from the integrated synergies in MixITS-Kit, rather than relying only on the isolated use of its components.

Our results suggest that our toolkit could offer shared vocabulary among designers, as participants found its language easy to communicate with and understand. The proposed design patterns further contribute to a shared vocabulary. As suggested by Alexander et al. (1977), design patterns externalize and document recurrent design decisions to foster a shared vocabulary, enabling collaborative and iterative refinement of solutions. Task 2 further supports this, with all participants successfully identifying the actor and target in the interactions.

7.2 MixITS class reflection

The extended 10-week course, compared to a short-term workshop, provided the instructional team with deeper insights into effectively teaching MixITS system design principles. This format revealed common challenges, knowledge gaps, and misconceptions faced by novices. A key hurdle was shifting from this technology-driven mindset to a user-centered approach. Students particularly struggled with selecting appropriate interaction modalities, effective task segmentation, considering multiple task completion paths, and anticipating potential user and AI errors.

Role-play exercises with the Wizard of Oz technique allowed participants to understand user needs and technology limitations of AI and mixed reality. These test exercises revealed user difficulties such as misunderstanding the system's capabilities, performing unintended actions, attempting to interact with non-existent features, and trying to communicate with the AI as if it was another human. This didactic intervention proved effective in changing students' mindsets from technology-centered to user-centered. As students' understanding of user needs grew, they started to propose features grounded in user behavior rather than solely technological feasibility. This evolution in approach marked a significant and desirable shift that was evident in the final project designs.

Our toolkit developed from these insights and can help designers avoid common technology-centric pitfalls such as overreliance on touch interfaces, voice commands without context awareness, or assuming constant internet connectivity. Overall, the toolkit encourages designers to transcend established AI interaction paradigms, moving beyond simple prompting or screen-based

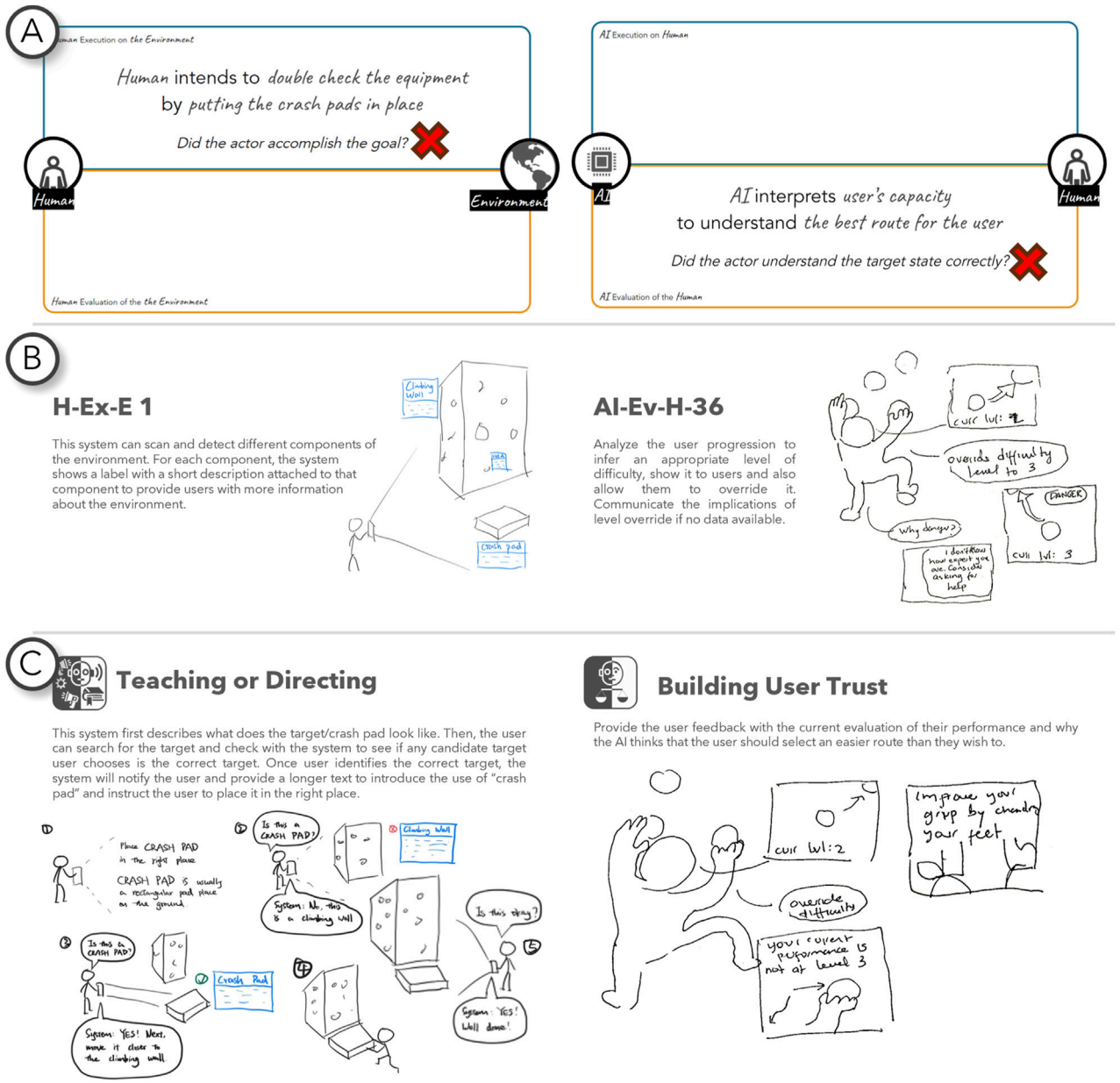

FIGURE 6
(A) In Task 1, participants used Interaction Canvases to identify a Gulf on fictional-user issues. **(B)** They then referred to Table 1 to find a problem that most closely matched their scenario and identified the applicable Design Pattern. **(C)** In Task 3, they revisited their initial solution through the lens of a Design Consideration. P6 used descriptive AR labels (H-Ex-E-1) with didactic instructions on the importance of crashpads and their use, considering "Teaching and Directing". After reflecting on "Build Trust", P10 added an AI explanation to the goal inferred by AI (AI-Ev-H-36).

interfaces towards a more holistic, embodied, and multi-modal approach essential for MixITS systems.

8 Limitations and future work

8.1 Curricular bias

We adopted an in-person classroom approach as a method for collecting instances of MixITS system designs and accompanying design processes. Among the limitations of this methodology choice are the particular perspectives on MixITS design arising from the choice of topics covered in class, selected reading material, the presentation approach of the class content in lectures and activities, proposed prototyping tools, and assignment requirements. Even though the teaching team made a continuous effort to foster discussions acknowledging multiple perspectives on the MixITS domain, we recognize that our teaching approach may have introduced its own biases. Moreover, we acknowledge that while the participants in our

formative study and evaluation represent early-career designers, engineers, and researchers who would use MixITS, future work should expand the MixITS-Kit with data collected from more experienced professionals in the industry.

8.2 Interactive MixITS-Kit

Using our Toolkit, novices in the MixITS domain were able to analyze interaction problems from a user perspective, propose solutions, and refine them within an hour, including the initial onboarding. However, our evaluation revealed areas for improvement in the Canvas and a need to lower the current threshold of the toolkit. Participants reported difficulties in identifying the gulfs (P4, P5, and P9) and suggested adding more detailed examples (P10) and clearer instructions (P4).

To address these issues, future work could involve developing an interactive version of MixITS-Kit, as suggested by P6. An integrated web platform that connects a browsable design pattern catalog with the interaction canvas would facilitate the use of strategies similar to those employed by P1 and P6, taking advantage of the synergy between toolkit components. This interactive version could also integrate data from user testing logs or the open-source MixITS database (Bohus et al., 2024) to ground designers' analyses, akin to (Castelo et al., 2023) but with a focus on interaction design rather than modeling.

8.3 Expanding MixITS-Kit

Our Design Pattern catalog is intended as a starting point, and we encourage others to replicate our methodology, refine our patterns, and add new ones. Re-purposing design patterns from related areas like mixed-initiative systems and intelligent tutoring systems could grow the list of MixITS design patterns. As consumer-grade MixITS products emerge, their inclusion as case studies will also help broaden the analysis pool, as seen in other design guidelines derives from web search, activity tracking, or recommendations (Amershi et al., 2019). In this work, we refrain from including MixITS solutions we implemented in the past or hypothetical ones not proposed by students. Consequently, some patterns presented in a gulf might also be relevant in another. For example, H-Ev-E-17 (Tool Operation) could easily be applied to the gulf of execution to instruct users on how to operate a tool. We encourage designers to consider the potential for re-purposing patterns across different gulfs, interpreting them from the perspective necessary for the design of their specific MixITS systems.

Our MixITS design patterns are derived from low-fidelity prototypes. Even though these designs are plausible, the course did not focus on the engineering challenges of such systems which have been explored in prior work (Bohus et al., 2021; Anderson et al., 1985; Andrist et al., 2019; Castelo et al., 2023). As the realtime task guidance field evolves, implementing these patterns in productive systems and evaluating their effectiveness with real users will be crucial to validate their practicality, along the lines of an evaluation of existing systems by Amershi et al. (2019).

Another promising direction for future work is to broaden our formative and evaluative scope to include experienced designers and industry professionals. Their insights would help us refine the toolkit to match the needs of experts; while also providing a valuable comparison to the findings we obtained from early-career designers.

9 Conclusion

In this work, we proposed an interaction design toolkit to support the design of AI systems for intelligent task support in mixed reality (MixITS). Drawing from MixITS prototypes developed during a 10-week graduate course on human-AI interaction, we derived six design considerations, an accompanying set of 36 design patterns, and an interaction canvas to help analyze and identify gulfs of execution and evaluation between the three MixITS entities of the user, the AI, and the environment. Our work aims to inspire and support the design and development of MixITS systems that blend the digital and physical worlds, creating situated and context-aware learning and task guidance experiences. Successfully tackling design challenges in the MixITS domain can play a significant role in ensuring that the transformative potential of AI is accessible to a wider range of users by augmenting the process of skill acquisition.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by University of California Santa Barbara Human Subjects Committee. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

AC: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review and editing. AA: Conceptualization, Data curation, Investigation, Methodology, Writing – original draft, Writing – review and editing. MS: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Investigation, Writing – original draft, Writing – review and editing, Data curation, Methodology, Validation.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. Arthur Caetano was

supported by the U.S. National Science Foundation (Early CAREER Award 2023 no. 2240133).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Correction note

A correction has been made to this article. Details can be found at: 10.3389/frvir.2025.1679670.

Generative AI statement

The authors declare that Gen AI was used in the creation of this manuscript. Editing: To draft alternative phrasing to some of authors' original writing. To summarize authors' original writing.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alexander, C., Ishikawa, S., Silverstein, M., Jacobson, M., Fiksdahl-King, I., and Shlomo, A. (1977). *engA pattern language: towns, buildings, construction. Center for Environmental Structure series*, 2. New York: Oxford University Press.

Allen, J. E., Guinn, C. I., and Horvtz, E. (1999). Mixed-initiative interaction. *IEEE Intelligent Syst. their Appl.* 14, 14–23. doi:10.1109/5254.796083

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., et al. (2019). "Guidelines for human-ai interaction," in *Proceedings of the 2019 chi conference on human factors in computing systems*, 1–13.

Anderson, F., Grossman, T., Matejka, J., and Fitzmaurice, G. (2013). "Youmove: enhancing movement training with an augmented reality mirror," in *Proceedings of the 26th annual ACM symposium on User interface software and technology*, 311–320.

Anderson, J. R., Boyle, C. F., and Reiser, B. J. (1985). Intelligent tutoring systems. *Science* 228, 456–462. doi:10.1126/science.228.4698.456

Andrist, S., Bohus, D., and Feniello, A. (2019). "Demonstrating a framework for rapid development of physically situated interactive systems," in 2019 14th ACM/IEEE international Conference on human-robot interaction (HRI) *(IEEE)*, 668.

Aponte, A., Caetano, A., Luo, Y., and Sra, M. (2024). "Grav: grasp volume data for the design of one-handed xr interfaces," in *Proceedings of the 2024 ACM designing interactive systems conference*, 151–167.

Apple (2024). Designing for visionos

Artizzu, V., Luyten, K., Ruiz, G. R., and Spano, L. D. (2024). Virgilites: multilevel feedforward for multimodal interaction in vr. *Proc. ACM Human-Computer Interact.* 8, 1–24. doi:10.1145/3658645

Ashtari, N., Bunt, A., McGrenere, J., Nebeling, M., and Chilana, P. K. (2020). "Creating augmented and virtual reality applications: current practices, challenges, and opportunities," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–13.

Bernal, G., Hidalgo, N., Russomanno, C., and Maes, P. (2022). "Galea: a physiological sensing system for behavioral research in virtual environments," in *2022 IEEE conference on virtual reality and 3D user interfaces (VR)* (IEEE), 66–76.

Bohus, D., Andrist, S., Feniello, A., Saw, N., Jalobeanu, M., Sweeney, P., et al. (2021). Platform for situated intelligence. *arXiv preprint arXiv:2103.15975*

Bohus, D., Andrist, S., Saw, N., Paradiso, A., Chakraborty, I., and Rad, M. (2024). "Sigma: an open-source interactive system for mixed-reality task assistance research – extended abstract," in *2024 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW)*, 889–890.

Borchers, J. O. (2000). "A pattern approach to interaction design," in *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*, 369–378.

Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qual. Res. Psychol.* 3, 77–101. doi:10.1191/1478088706qp063oa

Braun, V., and Clarke, V. (2024). A range of ways of approaching (reflexive) ta

Brown, J., and Burton, R. R. (1994). "Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Sci. 2, 155–192. doi:10.1207/s15516709cog0202_4

Burns, C., Dishman, E., Verplank, W., and Lassiter, B. (1994). "Actors, hairdos and videotape—informance design," in *Conference companion on Human factors in computing systems*, 119–120.

Byrne, R., Marshall, J., and Mueller, F. (2016). "Balance ninja: towards the design of digital vertigo games via galvanic vestibular stimulation," in *Proceedings of the 2016 annual symposium on computer-human interaction in play*, 159–170.

Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. *New Dir. Program Eval.* 1986, 67–77. doi:10.1002/ev.1434

Castelo, S., Rulff, J., McGowan, E., Steers, B., Wu, G., Chen, S., et al. (2023). Argus: visualization of ai-assisted task guidance in ar. *IEEE Trans. Vis. Comput. Graph.* 30, 1313–1323. doi:10.1109/tvcg.2023.3327396

Chatterjee, I., Pforte, T., Tng, A., Salemi Parizi, F., Chen, C., and Patel, S. (2022). "Ardw: an augmented reality workbench for printed circuit board debugging," in *Proceedings of the 35th annual ACM symposium on user interface software and technology*, 1–16.

Choi, H.-H., Van Merriënboer, J. J., and Paas, F. (2014). Effects of the physical environment on cognitive load and learning: towards a new model of cognitive load. *Educ. Psychol. Rev.* 26, 225–244. doi:10.1007/s10648-014-9262-6

Coppers, S., Luyten, K., Vanacken, D., Navarre, D., Palanque, P., and Gris, C. (2019). Fortunettes: feedforward about the future state of gui widgets. *Proc. ACM Human-Computer Interact.* 3, 1–20. doi:10.1145/3331162

Council, D. (2024). The double diamond

Curtis, D., Mizell, D., Gruenbaum, P., and Janin, A. (1999). "Several devils in the details: making an ar application work in the airplane factory," in *Proc. Int'l workshop augmented reality*, 47–60.

De Sá, M., and Churchill, E. (2012). "Mobile augmented reality: exploring design and prototyping techniques," in *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*, 221–230.

Ellenberg, M. O., Satkowski, M., Luo, W., and Dachselt, R. (2023). Spatiality and semantics-towards understanding content placement in mixed reality. *Ext. Abstr. 2023 CHI Conf. Hum. Factors Comput. Syst.*, 1–8. doi:10.1145/3544549.3585853

Elsden, C., Tallyn, E., and Nissen, B. (2020). "When do design workshops work (or not)?," in *Companion publication of the 2020 ACM designing interactive systems conference*, 245–250.

Erez, A., and Isen, A. M. (2002). The influence of positive affect on the components of expectancy motivation. *J. Appl. Psychol.* 87, 1055–1067. doi:10.1037/0021-9010.87.6.1055

Ericsson, K. A., and Simon, H. A. (1980). Verbal reports as data. *Psychol. Rev.* 87, 215–251. doi:10.1037/0033-295x.87.3.215

Eyck, A., Geerlings, K., Karimova, D., Meerbeek, B., Wang, L., Ijsselsteijn, W., et al. (2006). "Effect of a virtual coach on athletes' motivation," in *Persuasive technology: first international conference on persuasive technology for human well-being, PERSUASIVE 2006, eindhoven, The Netherlands, may 18-19, 2006. Proceedings 1* (Springer), 158–161.

Feiner, S., MacIntyre, B., and Seligmann, D. (1993). Knowledge-based augmented reality. *Commun. ACM* 36, 53–62. doi:10.1145/159544.159587

Feng, K. K., Coppock, M. J., and McDonald, D. W. (2023). "How do ux practitioners communicate ai as a design material? artifacts, conceptions, and propositions," in *Proceedings of the 2023 ACM designing interactive systems conference*, 2263–2280.

Flavell, J. H. (1976). "Metacognitive aspects of problem solving," in *The nature of intelligence* (London: Routledge), 231–236. doi:10.4324/9781032646527

Freitas, G., Pinho, M. S., Silveira, M. S., and Maurer, F. (2020). "A systematic review of rapid prototyping tools for augmented reality," in *2020 22nd symposium on virtual and augmented reality (SVR)* (IEEE), 199–209.

Haesler, S., Kim, K., Bruder, G., and Welch, G. (2018). "Seeing is believing: improving the perceived trust in visually embodied alexa in augmented reality," in *2018 IEEE international symposium on mixed and augmented reality adjunct (ISMAR-Adjunct)* (IEEE), 204–205.

Harp, S. F., and Mayer, R. E. (1997). The role of interest in learning from scientific text and illustrations: on the distinction between emotional interest and cognitive interest. *J. Educ. Psychol.* 89, 92–102. doi:10.1037//0022-0663.89.1.92

Henderson, S., and Feiner, S. (2010). Exploring the benefits of augmented reality documentation for maintenance and repair. *IEEE Trans. Vis. Comput. Graph.* 17, 1355–1368. doi:10.1109/tvcg.2010.245

Hirzle, T., Müller, F., Draxler, F., Schmitz, M., Knierim, P., and Hornbæk, K. (2023). "When xr and ai meet-a scoping review on extended reality and artificial intelligence," in *Proceedings of the 2023 CHI conference on human factors in computing systems*, 1–45.

Hornbæk, K., and Oulasvirta, A. (2017). "What is interaction?," in *Proceedings of the 2017 CHI conference on human factors in computing systems*, 5040–5052.

Horvitz, E. (1999). Principles of mixed-initiative user interfaces. *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.* 99, 159–166. doi:10.1145/302979.303030

Hutchins, E. L., Hollan, J. D., and Norman, D. A. (1985). Direct manipulation interfaces. *Human–computer Interact.* 1, 311–338. doi:10.1207/s15327051hci0104_2

Kelley, J. F. (1983). An empirical methodology for writing user-friendly natural language computer applications. *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 193–196. doi:10.1145/800045.801609

Koedinger, K. R., and Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educ. Psychol. Rev.* 19, 239–264. doi:10.1007/s10648-007-9049-0

Laurillard, D. (2013). *Teaching as a design science: building pedagogical patterns for learning and technology*. New York: Routledge. doi:10.4324/9780203125083

LaViola, J. (2017). *eng*3D user interfaces. 2nd edition edn. Addison-Wesley Professional.

Ledo, D., Houben, S., Vermeulen, J., Marquardt, N., Oehlberg, L., and Greenberg, S. (2018). "Evaluation strategies for hci toolkit research," in *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–17.

Leiva, G., Nguyen, C., Kazi, R. H., and Asente, P. (2020). "Pronto: rapid augmented reality video prototyping using sketches and enaction," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–13.

Li, T., Vorvoreanu, M., DeBellis, D., and Amershi, S. (2023). Assessing human-ai interaction early through factorial surveys: a study on the guidelines for human-ai interaction. *ACM Trans. Computer-Human Interact.* 30, 1–45. doi:10.1145/3511605

Liu, X. B., Li, J. N., Kim, D., Chen, X., and Du, R. (2024). Human i/o: towards a unified approach to detecting situational impairments. *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 1–18. doi:10.1145/3613904.3642065

Lupetti, M. L., and Murray-Rust, D. (2024). "(un) making ai magic: a design taxonomy," in *Proceedings of the CHI conference on human factors in computing systems*, 1–21.

Mandic, S., Tracy, R., and Sra, M. (2023). "Arfit: pose-based exercise feedback with mobile ar," in *Proceedings of the 2023 ACM symposium on spatial user interaction*, 1–3.

Maxwell, J. A. (2010). Using numbers in qualitative research. *Qual. Inq.* 16, 475–482. doi:10.1177/1077800410364740

Meta (2024). Designing for mixed reality

Microsoft (2024). Thinking differently for mixed reality

Milgram, P., Takemura, H., Utsumi, A., and Kishino, F. (1995). Augmented reality: a class of displays on the reality-virtuality continuum. *Telemanipulator telepresence Technol. (Spie)* 2351, 282–292. doi:10.1117/12.197321

Muresan, A., McIntosh, J., and Hornbæk, K. (2023). Using feedforward to reveal interaction possibilities in virtual reality. *ACM Trans. Computer-Human Interact.* 30, 1–47. doi:10.1145/3603623

Nith, R., Ho, Y., and Lopes, P. (2024). "Splitbody: reducing mental workload while multitasking via muscle stimulation," in *Proceedings of the CHI conference on human factors in computing systems*, 1–11.

Norman, D. A. (1986). User-centered system design: new perspectives on human–computer interaction

Norman, D. (2013). *The design of everyday things: revised and expanded edition*. Basic Books.

Norman, D. A. (1981). Categorization of action slips. *Psychol. Rev.* 88, 1–15. doi:10.1037//0033-295x.88.1.1

Oulasvirta, A., Kurvinen, E., and Kankainen, T. (2003). Understanding contexts by being there: case studies in bodystorming. *Personal ubiquitous Comput.* 7, 125–134. doi:10.1007/s00779-003-0238-7

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* 35, 27730–27744.

Polit, D. F., and Beck, C. T. (2010). Generalization in quantitative and qualitative research: myths and strategies. *Int. J. Nurs. Stud.* 47, 1451–1458. doi:10.1016/j.ijnurstu.2010.06.004

Rauh, S. F., Bogdan, C., Meixner, G., and Matviienko, A. (2024). "Navigating the virtuality-reality clash: reflection and design patterns for industrial mixed reality applications," in *Proceedings of the 2024 ACM designing interactive systems conference*, 2247–2266.

Retalis, S., Georgiakakis, P., and Dimitriadis, Y. (2006). Eliciting design patterns for e-learning systems. *Comput. Sci. Educ.* 16, 105–118. doi:10.1080/08993400600773323

Rheault, B., Arya, S., Vyas, A., Wang, J., Peddi, R., Bendall, B., et al. (2024). "Predictive task guidance with artificial intelligence in augmented reality," in *2024 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW)* (IEEE), 973–974.

Rigby, L., Wünsche, B. C., and Shaw, A. (2020). "piarno-an augmented reality piano tutor," in *Proceedings of the 32nd Australian conference on human-computer interaction*, 481–491.

Schwerdtfeger, B., Frimor, T., Pustka, D., and Klinker, G. (2006). "Mobile information presentation schemes for supra-adaptive logistics applications," in *Advances in artificial reality and tele-existence: 16th international conference on artificial reality and telexistence, ICAT 2006, hangzhou, China, november 29-december 1, 2006* (Proceedings: Springer), 998–1007.

Shahu, A., Dorfbauer, S., Wintersberger, P., and Michahelles, F. (2023). "Skillab - a multimodal augmented reality environment for learning manual tasks," in *Human-computer interaction – interact 2023*. Editors J. Abdelnour Nocera, M. Kristín Lárusdóttir, H. Petrie, A. Piccinno, and M. Winckler (Cham: Springer Nature Switzerland), 588–607.

Shappell, S. A., and Wiegmann, D. A. (2000). The human factors analysis and classification system–hfacs

Sharma, A., Hedderich, M. A., Bhardwaj, D., Fruchard, B., McIntosh, J., Nittala, A. S., et al. (2021). "Solofinger: robust microgestures while grasping everyday objects," in *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–15.

Sosnowski, T., Stoev, T., Kirste, T., and Yordanova, K. (2023). "Challenges in modelling cooking task execution for user assistance," in *Proceedings of the 8th international workshop on sensor-based activity recognition and artificial intelligence*, 1–4.

Sra, M., Xu, X., and Maes, P. (2017). "Galvr: a novel collaboration interface using gvs," in *Proceedings of the 23rd ACM symposium on virtual reality software and technology*, 1–2.

St. Amant, R., and Cohen, P. R. (1997). "Interaction with a mixed-initiative system for exploratory data analysis," in *Proceedings of the 2nd international conference on Intelligent user interfaces*, 15–22.

Svanaes, D., and Seland, G. (2004). "Putting the users center stage: role playing and low-fi prototyping enable end users to design mobile systems," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 479–486.

Tang, A., Owen, C., Biocca, F., and Mou, W. (2003). Comparative effectiveness of augmented reality in object assembly. *Proc. SIGCHI Conf. Hum. factors Comput. Syst.*, 73–80. doi:10.1145/642611.642626

Terry, M., Kulkarni, C., Wattenberg, M., Dixon, L., and Morris, M. R. (2023). Ai alignment in the design of interactive ai: specification alignment, process alignment, and evaluation support. *arXiv Prepr. arXiv:2311.00710*.

Thomas, P. C., and David, W. (1992). Augmented reality: an application of heads-up display technology to manual manufacturing processes. *Hawaii Int. Conf. Syst. Sci. ACM SIGCHI Bull.* 2, 659–669. doi:10.1109/HICSS.1992.183317

Tiator, M., Geiger, C., Dewitz, B., Fischer, B., Gerhardt, L., Nowottnik, D., et al. (2018). "Venga! climbing in mixed reality," in *Proceedings of the first superhuman sports design challenge: first international symposium on amplifying capabilities and competing in mixed realities*, 1–8.

Tresselt, M., and Mayzner, M. (1960). A study of incidental learning. *J. Psychol.* 50, 339–347. doi:10.1080/00223980.1960.9916451

Ultraleap (2024). Xr design guidelines

van Asselen, M., Fritschy, E., and Postma, A. (2006). The influence of intentional and incidental learning on acquiring spatial knowledge during navigation. *Psychol. Res.* 70, 151–156. doi:10.1007/s00426-004-0199-0

Vermeulen, J., Luyten, K., van den Hoven, E., and Coninx, K. (2013). Crossing the bridge over norman's gulf of execution: revealing feedforward's true identity. *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 1931–1940. doi:10.1145/2470654.2466255

Wang, P., Bai, X., Billinghurst, M., Zhang, S., Zhang, X., Wang, S., et al. (2021). Ar/mr remote collaboration on physical tasks: a review. *Robotics Computer-Integrated Manuf.* 72, 102071. doi:10.1016/j.rcim.2020.102071

Weaver, K. A., Baumann, H., Starner, T., Iben, H., and Lawo, M. (2010). "An empirical task analysis of warehouse order picking using head-mounted displays," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 1695–1704.

Winters, N., and Mor, Y. (2009). Dealing with abstraction: case study generalisation as a method for eliciting design patterns. *Comput. Hum. Behav.* 25, 1079–1088. doi:10.1016/j.chb.2009.01.007

Wu, G., Qian, J., Castelo Quispe, S., Chen, S., Rulff, J., and Silva, C. (2024). "Artist: automated text simplification for task guidance in augmented reality," in *Proceedings of the CHI conference on human factors in computing systems*, 1–24.

Xu, X., Yu, A., Jonker, T. R., Todi, K., Lu, F., Qian, X., et al. (2023). "Xair: a framework of explainable ai in augmented reality," in *Proceedings of the 2023 CHI conference on human factors in computing systems*, 1–30.

Yang, Q., Steinfeld, A., Rosé, C., and Zimmerman, J. (2020). "Re-examining whether, why, and how human-ai interaction is uniquely difficult to design," in *Proceedings of the 2020 chi conference on human factors in computing systems*, 1–13.

Yildirim, N., Kass, A., Tung, T., Upton, C., Costello, D., Giusti, R., et al. (2022). "How experienced designers of enterprise applications engage ai as a design material," in *Proceedings of the 2022 CHI conference on human factors in computing systems*, 1–13.

Yoo, D. W., Tarashiyoun, H., and Moghaddam, M. (2023). "Modeling gaze behavior for real-time estimation of visual attention and expertise level in augmented reality," in *2023 IEEE international symposium on mixed and augmented reality adjunct (ISMAR-Adjunct)* (IEEE), 487–492.