Should the AI Speak First? Evaluating Proactive vs. Reactive Facilitation in Mixed-Reality Medical Training

Duo Wang
University of Illinois Urbana Champaign
Champaign, Illinois, USA
dw36@illinois.edu

Wenxuan Song
University of Illinois Urbana Champaign
Champaign, Illinois, USA
js129@illinois.edu

Jianwei Ni
University of Illinois Urbana Champaign
Champaign, Illinois, USA
ni12@illinois.edu

Qingxiao Zheng
University at Buffalo
Buffalo, New York, USA
qingxiao@buffalo.edu

Yuhan Zhou
University of Illinois Urbana Champaign
Champaign, Illinois, USA
yuhan15@illinois.edu

Kyrian Liang
University of Illinois Urbana Champaign
Champaign, Illinois, USA
jingxil2@illinois.edu

Mike Yao
University of Illinois Urbana Champaign
Champaign, Illinois, USA
Mzyao@illinois.edu

Caroline Cao
University of Ottawa
Ottawa, Ontario, Canada
caroline.cao@uottawa.ca

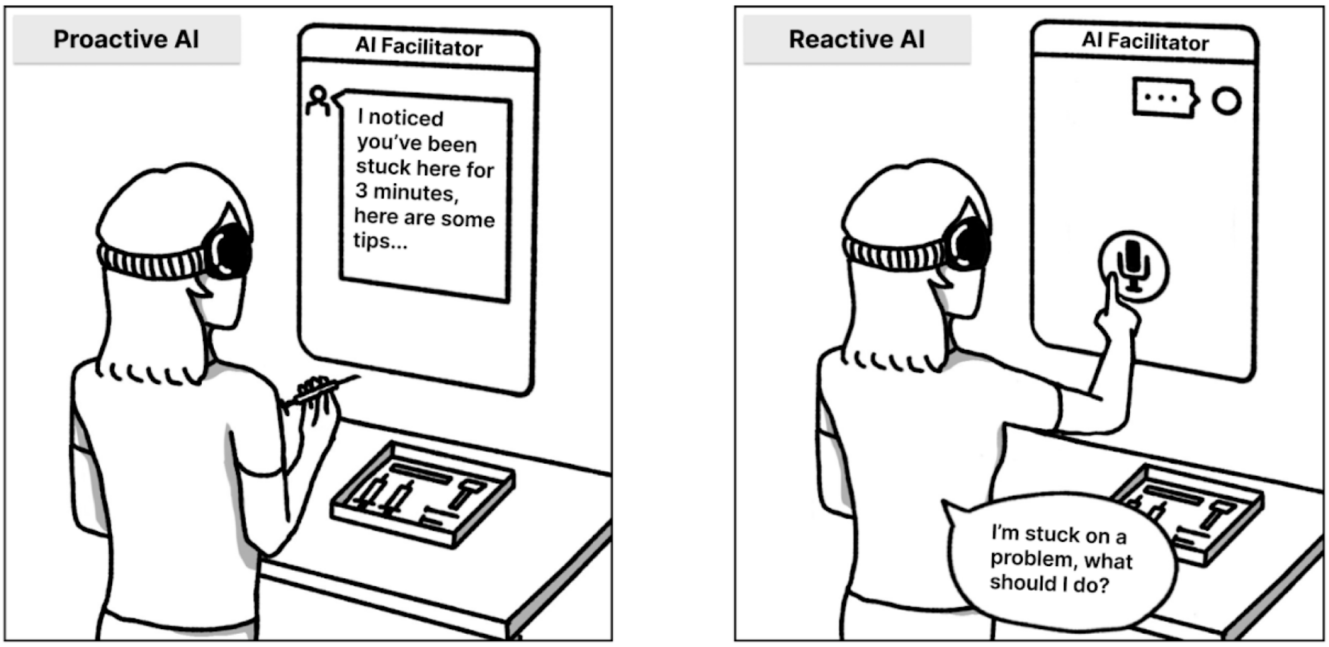


Figure 1: Conceptual illustration of the two AI facilitator conditions examined in this study. Both conditions operate within the same XR training environment and differ only in how the AI allocates initiative. In the Proactive AI + XR condition (left), the AI monitors learner progress and offers guidance when it detects potential difficulty. In the Reactive AI + XR condition (right), the AI remains silent until the learner explicitly requests support.

ABSTRACT

As AI support tools become more common in immersive medical training, designers face a critical interaction-design question: When should an AI facilitator take initiative, and when should it wait for the learner? To investigate this design tension, we compared two versions of an AI facilitator in a mixed-reality (XR) lumbar puncture simulator training conditions: one in which the AI proactively initiated guidance and encouragement, and another in which the AI responded only when prompted. Using a mixed-methods approach, we examined how medical students (n=22) engaged with, interpreted, and reacted to these two facilitation styles. We found no significant differences in learning outcomes, interaction frequency, or overall experience ratings. However, interviews and behavioral analyses revealed nuanced differences in how learners perceived AI interventions across distinct task phases. AI-initiated support was seen as helpful in some moments and disruptive in others, depending on task phase, cognitive load, and personal preferences. Based on these findings, we contribute a boundary framework which offers actionable design guidance for calibrating AI proactivity in immersive training systems, and extends HCI research on proactive agents and human–AI collaboration within high-cognitive-load environments.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

Medical Simulation, AI Facilitator, Mixed Reality

ACM Reference Format:

Duo Wang, Wenxuan Song, Jianwei Ni, Qingxiao Zheng, Yuhan Zhou, Kyrian Liang, Mike Yao, and Caroline Cao. 2026. Should the AI Speak First? Evaluating Proactive vs. Reactive Facilitation in Mixed-Reality Medical Training. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3772318.3791910

1 INTRODUCTION

Simulation is widely used in medical education to help learners practice procedures like lumbar puncture in safe, controlled environments [22, 25, 92]. High-quality facilitation is a critical factor to simulation success. In medical contexts, facilitators guide technical steps, correct errors, and help learners interpret cues and outcomes as they perform [37, 65]. However, as simulation programs grow, the reliance on trained facilitators creates a persistent bottleneck: access becomes limited, feedback inconsistent, and sessions difficult to scale [100, 104].

Mixed Reality (MR) technologies are increasingly used to enhance medical simulation by layering virtual content onto physical tools and manikins [91, 92]. These immersive environments can help standardize training, support self-paced learning, and reduce reliance on fully staffed simulation rooms [24, 47]. However, MR alone does not meet the need for instructional guidance. Without real-time support or feedback, learners may develop misconceptions, miss critical steps, or struggle to interpret complex procedural cues, especially in simulations involving tactile skill execution [4].

To address the shortage of expert facilitators in medical simulation, researchers and designers have increasingly explored the use of AI to deliver procedural guidance, feedback, and encouragement during training [24, 111]. Prior work comparing AI-enhanced XR systems with XR alone consistently demonstrates that integrating AI improves learning outcomes, efficiency, and engagement by providing more personalized, responsive, and scalable support [27, 33, 93, 103]. As these systems become more capable, a key emerging challenge is determining how AI facilitators should behave during simulation [28]. *Should the AI wait for learners to request help, or proactively step in with hints, check-ins, or encouragement?* AI-initiated support can keep learners on track and reduce confusion [7, 36], yet in immersive simulations—where learners are physically engaged and cognitively absorbed—unsolicited interventions may feel intrusive or reduce learner agency [2]. These tensions foreground a broader design question: *when AI is embedded in simulation systems, how much initiative should it take?* While intelligent tutoring systems have examined related issues of initiative and timing [34, 44, 98, 107], little is known about how these dynamics unfold in medical simulation training conducted within embodied, immersive environments.

To explore this design tension, we developed an immersive simulation that supports procedural training with an embedded AI facilitator. We implemented two contrasting strategies: one in which the AI offers help proactively, and another in which it only responds to learner prompts. This comparison enabled us to examine how the timing and initiative of AI support impact learners' experience, behavior, and perceived effectiveness in medical simulations conducted within immersive environments.

Guided by this comparison, we asked the following research questions:

- **RQ1:** How does the AI facilitator's mode of support affect learners' experience and perceived learning outcomes in immersive medical simulation?
- **RQ2:** How do learners interact differently with AI that provides proactive support versus AI that responds only when prompted?
- **RQ3:** What design opportunities and boundaries emerge for implementing AI-guided support in immersive simulation-based training?

Through this study, we contribute empirical insights and design implications for AI-facilitated immersive simulation. Our work advances HCI research on *AI proactivity* by examining how AI should interact with users in embodied, spatial, and high-stakes procedural contexts. Specifically, (1) We examine how an AI facilitator should behave once integrated into an XR procedural training environment—a question not addressed in prior work that primarily contrasts AI-supported XR with XR-only systems. Through a controlled study comparing proactive and reactive AI facilitation, we move beyond the question of *"Does AI help?"* toward understanding how initiative, timing, and situational relevance shape learner

Please use nonacm option or ACM Engage class to enable CC licenses
This work is licensed under a Creative Commons Attribution 4.0 International License.
CHI '26, April 13–17, 2026, Barcelona, Spain
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3791910

experience in embodied, hands-on tasks; (2) We contribute a mixed-methods analysis, including behavioral video coding, quantitative measures, stimulated recall interviews, and learner–AI interaction journeys, to reveal how AI interventions are interpreted differently across the phases of a procedural workflow; And (3) We propose a boundary framework of Initiative, Temporal, and Situational Boundaries, mapped to four key moments in immersive learning. This framework synthesizes insights from the proactive AI literature with the unique demands of XR procedural training and offers actionable design guidance for calibrating AI proactivity to support learners without disrupting autonomy or task flow.

2 RELATED WORK

2.1 Facilitators in Medical Training

Medical training is a complex and high-stakes learning process that requires learners to master both cognitive knowledge and hands-on procedural skills [88]. Unlike traditional classroom education, medical training frequently incorporates simulation-based practice as a bridge between theoretical instruction and real-world patient care to enable learners to engage in experiential learning without endangering patients [24, 78]. These simulations replicate authentic clinical environments and offer opportunities to practice diagnostic, surgical, and emergency response skills within a structured and safe setting that reinforces the development of contextual competence [3]. Research suggests that simulation-based training contributes to increased procedural accuracy, enhanced clinical reasoning, and greater readiness for independent care delivery [24, 79, 84]. However, simulation-based training introduces several challenges. Learners often face difficulty navigating unfamiliar equipment or procedural steps, especially under stressful or high-pressure conditions, which can inhibit performance and learning outcomes [90]. Without the presence of real-time guidance or timely feedback, trainees may fail to recognize critical errors or reinforce correct behaviors, which limits their confidence and self-efficacy during independent clinical tasks [11].

To address the limitations of traditional, passive learning approaches, many training systems now incorporate facilitators who guide learners through complex procedures, offer clarification, and provide motivational support [64, 95]. In the context of medical training, such facilitators play a crucial role in supporting learners' progression across canonical and contextual layers of competence [88, 97]. By intervening during key moments, correcting misconceptions, and reinforcing correct techniques, facilitators reduce learners' cognitive load and emotional stress during clinical tasks [26, 86]. However, facilitator-led training faces practical and scalability challenges. Human facilitators may not be consistently available, especially when training large cohorts or integrating advanced immersive modalities such as virtual or mixed reality, which require additional supervision resources [15, 50]. Conversely, digital facilitators often depend on rigid, pre-scripted interactions, which can fall short of responding to learners' unique prior knowledge, emotional states, or evolving needs during practice [41, 46]. This highlights the need for more adaptive facilitator systems that can respond in real time to individual learner behaviors and situational demands.

2.2 How AI Amplifies XR Training: Insights From Prior Literature

To address the limitations of time-constrained human supervision in medical simulation, recent innovations have turned toward AI facilitators that offer scalable, responsive support during procedural training [23, 99, 106]. We use the term AI facilitators broadly to include intelligent tutoring systems, adaptive pedagogical agents, and agent-based guidance mechanisms—systems that algorithmically interpret learner behavior and deliver tailored instructional support. These agents leverage capabilities such as behavior tracking, natural language understanding, and adaptive feedback to respond dynamically to learners' actions in real time [39, 67, 101]. Rather than relying on static, one-size-fits-all prompts, AI agents can use multimodal cues such as gaze, hesitation, or interaction speed to identify when learners are struggling and intervene with adaptive, context-sensitive feedback [21]. This enables consistent instructional support across diverse training scenarios, even in the absence of a human instructor [16, 89]. Such systems have shown promise in domains such as intelligent tutoring and language learning, where responsiveness and personalization enhance learning outcomes [44, 61]. In medical education—where procedural accuracy and timing are critical—AI agents offer a compelling mechanism for guiding learners through complex workflows while reducing the burden on faculty resources [23].

The integration of extended reality (XR) platforms further amplifies the effectiveness of these AI facilitators by situating them within immersive, spatial learning environments [108]. Unlike traditional interfaces, XR allows AI agents to share the same three-dimensional workspace as learners [102]. This shared spatial context enables fine-grained inference of learner intent from gaze, hand motion, and tool usage [53], allowing AI facilitators to provide visual overlays, spatial audio, or gestural cues directly within the learner's field of view [8]. These modalities are especially valuable in medical training, where timely, well-situated feedback is essential [76]. Such embodied feedback channels align with the long-standing strengths of intelligent tutoring and pedagogical agent systems, but extend them into immersive, real-time spatial contexts. Thus, AI and XR together form a powerful foundation for learner support: AI provides adaptivity and intelligence, while XR provides embodied, contextualized interaction channels.

Given this complementary relationship, a growing body of research has directly compared XR-only systems with XR enhanced by intelligent tutoring or agent-based guidance. Across domains, these studies consistently show that adding AI-driven facilitation (e.g., adaptive intelligent tutoring, embodied agents, or intelligent AR guidance) to XR yields measurable improvements in learning outcomes, task performance, and learner engagement [103]. In an adaptive AR intelligent tutoring system for geometry learning, students using the intelligent agent (which adjusted task difficulty based on prior performance) achieved significantly higher learning gains than those using the same AR interface without the tutoring component (7.47 vs 6.83) [93]. Similarly, in AR-based motherboard assembly training, adding an intelligent tutoring module that delivered adaptive hints led to a 25% improvement in test scores and a 30% increase in task efficiency compared to an AR-only condition

[103]. Comparable findings emerge in virtual reality learning environments. Studies of VR pedagogical agents show that adding an embodied agent who provides explanations, motivational support, and context-sensitive feedback significantly improves learners' engagement and performance relative to VR-only conditions with static instructions [33]. More recent work in industrial assembly tasks demonstrates that "intelligent AR" systems featuring scene understanding, automatic instruction triggering, and adaptive next-step guidance enable unskilled workers to complete tasks more quickly (35%) and more accurately (43%) than when using identical AR hardware without the intelligent components [27]. Collectively, these findings demonstrate that XR paired with intelligent facilitation reliably outperforms XR alone.

2.3 Proactive AI Research in HCI

Across HCI, proactive agents are defined by their ability to anticipate user needs and initiate actions without explicit prompting [6, 18, 19, 56, 58]. Rather than waiting for commands, proactive systems infer intentions, detect breakdowns, and intervene at opportune moments to support task performance [52, 54]. They can identify hesitation, tool misplacement, or missed procedural steps and provide timely guidance without requiring user prompts [6, 49]. For example, a system takes initiative by modeling user state and stepping in before breakdowns occur, demonstrating how proactive intervention can increase efficiency and accuracy in interactive tasks [6]. In collaborative or educational settings, proactive agents have been shown to improve usability, task efficiency, and trust by offering context-sensitive guidance at key moments [13, 58, 110]. In contrast, passive agents take a nonintrusive approach, helping users maintain focus without interrupting task flow. For instance, prior work finds that passive AI was preferred in collaborative meetings because it supported reflection while preserving conversational rhythm [14]. These differing strategies highlight a longstanding tension in HCI: systems must balance initiative with the risk of interrupting or overriding user autonomy.

Recent AI systems highlight both the benefits and risks of AI proactivity aiming for tasks across different domains, such as health care [74, 81, 105], navigation [70], and laboratory education [82]. In software development, proactive programming assistants such as CodingGenie demonstrate that AI-initiated suggestions can significantly improve developer efficiency, though poorly timed interventions may disrupt workflow unless accompanied by presence indicators or contextual cues [109]. Additional studies on proactive programming support report similar patterns: proactive assistance increases efficiency but can introduce workflow interruptions when triggers are mistimed, underscoring the need for context-awareness in suggestion delivery [13, 73]. Proactive conversational agents have been shown to improve usability, trust, and goal achievement through inner-state reasoning [57], intention-based recommendation [55], and multimodal interaction modeling [13, 94], but emphasizing that initiative must be delivered at moments that align with user state and situational demands. Work in proactive digital companions further demonstrates reduced task load and increased satisfaction when initiative is aligned with user activity and attentional state [72]. In the robotic field, timely assistance can be helpful, but if an AI interrupts too often or at the wrong moment, it may frustrate learners or make them feel micromanaged [48]. Similar benefits appear in wellbeing support, where proactive voice assistants such as WorkFit facilitate healthier work habits through anticipatory nudges [1]. These systems, along with frameworks such as Proactive Agent [59] and surveys on proactive AI [18], collectively demonstrate that well-timed initiative can enhance task flow, reduce cognitive load, and improve satisfaction, but only when calibrated to user context, autonomy, and attentional state.

While proactive AI has been widely explored in screen-based and dialog-driven systems, mixed reality (XR) procedural training introduces interactional challenges that existing frameworks do not fully address. In XR, users are engaged in embodied, spatially distributed, and sequential workflows that require continuous coordination of gaze, posture, hand movement, and tool manipulation [51]. This creates a need to carefully balance the level of initiative the AI takes with the learner's sense of autonomy and control [45, 66]. Most XR assistance today remains passive because defining user intention is difficult [54]. One challenge is that understanding user intention relies not only on explicit cues (e.g., verbal statements or signals) but also, significantly, on implicit non-verbal cues and body postures [43]. Successfully decomposing and reasoning with implicit cues improves the chances of intention labeling. Recent proactive AR systems demonstrate the promise and complexity of bringing initiative-taking AI into these environments.

For example, YETI enables AR agents to detect context changes through multimodal signals and proactively intervene during hands-on tasks, outperforming reactive baselines in timeliness and support effectiveness [6]. Similarly, Sensible Agent reduces cognitive load by adapting what to suggest and how to deliver it through context-aware visual, audio, or gestural cues, offering proactive help without compromising social unobtrusiveness [52]. Satori further advances proactive AR by using belief–desire–intention modeling and multimodal LLMs to anticipate user goals and provide expert-level guidance across everyday manipulation tasks [54]. Beyond AR, research in virtual learning environments shows that proactive features can increase motivation and engagement but must be carefully calibrated to avoid disrupting immersion or autonomy [71].

Together, these findings suggest proactive AI benefits are highly context dependent and must be carefully calibrated to avoid disruption while preserving learner autonomy [45]. This sensitivity highlights the need for principled design guidance for how proactive AI should behave within the XR environment. Thus, we aim to propose a framework of three design boundaries mapped across four key training moments, providing actionable guidance for calibrating AI proactivity in the XR environment.

3 METHOD

3.1 System Overview

Lumbar puncture (LP) is a common but technically demanding procedure used to collect cerebrospinal fluid (CSF) for diagnostic and therapeutic purposes. We selected LP because it captures key challenges common to many procedural skills in medical training. It is a high-stakes procedure where mistakes can lead to serious patient harm, and learners have few opportunities for supervised practice [96]. The task requires fine-grained spatial judgment, adherence

to stepwise protocols, and timely corrective feedback—conditions that are often difficult to achieve with limited faculty availability and manikin-based training. These characteristics make lumbar puncture a representative case for examining how immersive AI systems can extend to other procedural domains with similar training constraints.

To support immersive lumbar puncture training, we developed a mixed reality (MR) system using the Apple Vision Pro headset that layered spatial guidance and AI support onto a real-world procedural environment. Medical trainees wore the headset to interact with a physical lumbar puncture manikin and a real instrument tray, which enables authentic hands-on engagement. The MR system displayed a transparent 3D spine model aligned with the manikin, along with visual cues, step indicators, and AI messages that responded to trainee actions. Using the Apple Vision Pro's built-in sensors, learners controlled the interface through natural gaze and hand gestures. This allowed them to perform physical tasks while seamlessly receiving digital support embedded in their field of view.

The system used the Apple Vision Pro's built-in eye tracking to monitor where learners focused their attention and hand tracking to detect when they approached the correct location for key procedural steps. For example, when a learner scanned the mannequin's lower back and moved their hand toward the appropriate palpation region, the system confirmed successful targeting. It triggered a short 3D animation overlay to reinforce the action.

The lumbar puncture tutorial is structured into eight sequential steps. Each step presents training materials in multiple formats, including text, voiceover, and 3D animation. Learners use a control panel to navigate between steps and access the instructional content. After reviewing the material, they can immediately practice the corresponding procedure on the manikin with real instruments while keeping the headset on. Once they feel comfortable with a given step, they can advance to the next stage of the tutorial.

To enhance this training experience, we designed two distinct AI Facilitators that varied in their level of initiative: a Proactive AI Facilitator and a Reactive AI Facilitator. These two modes represented contrasting approaches to AI engagement, allowing us to explore how differences in timing, learner autonomy, and AI-initiated guidance influence the overall learning experience in procedural simulation. Both AI modes were powered by a Gemini Flash language model, which generated context-aware responses based on real-time learner progress. Speech-to-text input was processed locally on the Apple Vision Pro for real-time responsiveness, while AI responses were synthesized using a server-side text-to-speech engine and streamed back to the headset with minimal delay.

3.1.1 Reactive AI Facilitator. The Reactive AI Facilitator provided support only when learners actively requested it. Instead of offering unsolicited help, the facilitator remained in the background, allowing participants to stay focused on the procedure. Learners could either tap the microphone icon to ask a question aloud or type their question into the right-side panel integrated into the mixed reality interface. In response, the AI offered guidance related to the learner's current task, such as explanations, tips, or step-by-step suggestions. These responses were presented in a clear and focused manner, often accompanied by visual cues to help learners follow along. This setup gave learners full control over when and how to engage with the AI, allowing them to receive help at their own pace without interrupting the flow of their training.

3.1.2 Proactive AI Facilitator. The Proactive AI Facilitator was designed to mirror the behavior of a mindful instructor who observes learner performance and offers timely guidance without being asked. Rather than waiting for explicit questions, this version continuously observed learners as they progressed through the lumbar puncture procedure. It used data such as how long learners spent on each step, where they looked, and whether their hands approached the correct location to decide when to intervene with support, reminders, or encouragement.

To ensure the AI remained helpful while respecting learner independence, we selected three types of behavioral triggers that shaped its timing and responses. Our selection of the three proactive behaviors—check-ins, reminders, and encouragement—was grounded in a combination of empirical observation, expert consultation, and evidence from prior literature. First, we analyzed several clinical lumbar puncture training sessions and noted that human facilitators consistently supported learners in three ways: re-engaging them when they hesitated, preventing them from skipping essential steps, and offering brief motivational comments to maintain confidence. These patterns shaped the initial feature set. Second, we held design discussions with clinical facilitators, who helped us identify which behaviors were both pedagogically valuable and technically feasible in an MR environment. Third, our review of work on mixed-initiative tutoring and coaching systems showed that similar forms of proactive assistance have been shown to benefit learners without undermining autonomy.

(1) **Check-in – Extended time on a step.** One important role of AI facilitators is to help learners re-engage when they show signs of hesitation or difficulty. Prior work shows that timely interventions can sustain engagement and help learners reorient when they struggle [5, 30, 32, 38, 60, 62, 68, 83]. Building on this insight, our system provided a gentle check-in when a learner spent more time than expected on a task, such as locating anatomical landmarks or inserting the needle. For example, the AI might say, "It seems you've been at this step for a while, so let's break it down."
(2) **Reminder - Quick progression through a step.** AI facilitators are also effective when they prevent learners from rushing through important steps. Studies in education, healthcare, and training emphasize that reminders can help learners slow down and attend to details that might otherwise be overlooked [32, 38, 60, 62, 68, 83]. In line with this, our system issued a reminder when a learner rushed through an action such as checking positioning or injecting an anesthetic. An example would be, "Are you sure you have practiced enough to continue?"
(3) **Encouragement - Completion on a step.** Beyond corrections, AI facilitators can play a motivational role by recognizing progress and reinforcing confidence. Research shows that brief, positive acknowledgments can enhance motivation and improve learning outcomes [12, 32, 38, 60, 62, 68]. Following this principle, our system provided short encouragement when a learner completed a step within a healthy

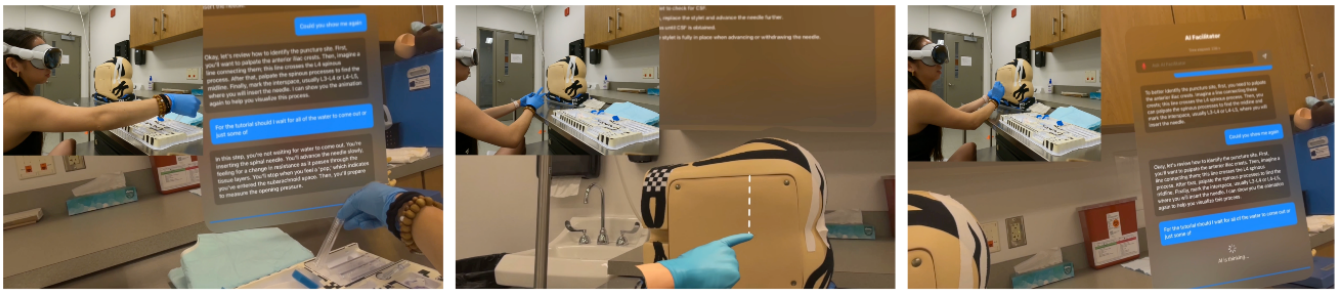


Figure 2: (a) participant interacting with real-world tools and manikin using the Apple Vision Pro headset, (b) 3D animation overlay guiding step-by-step spinal landmark identification, and (c) AI facilitator providing context-aware responses during the simulation.

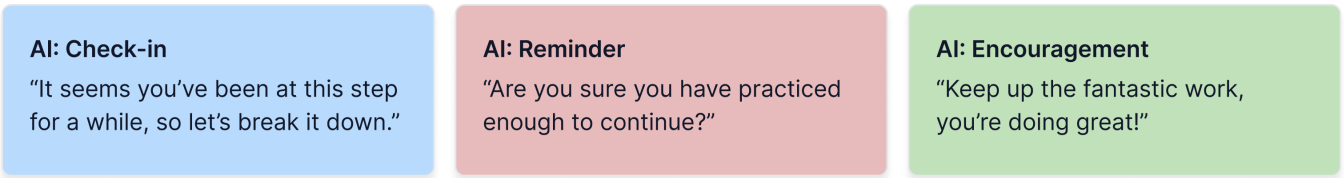


Figure 3: Conversation samples: three core features of proactive AI facilitator

time window, such as "Keep up the fantastic work, you're doing great!"

To identify these moments, we explored timing thresholds for each step in the procedure rather than setting them as a rigid rubric. These values were informed by pilot study observations and refined with feedback from Neurology Residents to capture patterns of hesitation, rushed behavior, and productive pacing. As shown in Table 1, the ranges provide a working reference for when the AI might offer a check-in, reminder, or encouragement. They represent a developmental stage of the system, allowing us to investigate how proactive facilitation can shape learning dynamics, rather than a finalized set of evaluation criteria.

3.2 User Study

To investigate how different levels of AI initiative influence learning experiences in immersive procedural training, we conducted a between-subjects study comparing two distinct AI conditions. Our primary goal was to understand how these AI designs shaped learner behavior, perceptions of support, and perceived boundaries of appropriate AI involvement over the course of the training experience.

3.2.1 Participants. Twenty-two medical students participated in the study. The majority were first-year students (n = 20), with 2 participants in their third year. All participants were enrolled in medical training programs and had completed coursework related to human anatomy. However, none had prior experience performing a lumbar puncture or using the extended reality training system. To ensure balanced exposure to the AI designs, participants were randomly assigned to one of the two experimental conditions: the Proactive AI Facilitator group (G1: n = 11) or the Reactive AI Facilitator group (G2: n = 11).

3.2.2 Study Procedure. Each session lasted approximately 1 hour and 30 minutes and took place in a dedicated extended reality (XR) training lab. All participants were recruited voluntarily through an IRB-approved flyer and provided informed consent before the session. To capture learner behavior and interaction dynamics, each session was recorded from two perspectives: an external third-person view using a GoPro camera and a first-person view using the Apple Vision Pro headset. This dual recording setup enabled a comprehensive analysis of gaze direction, physical actions, and real-time engagement with the AI support.

The session began with a 15-minute orientation phase where participants signed the consent form and completed a pre-survey. They were then introduced to the Vision Pro headset and practiced using its core features, including gaze-based selection, gesture control, and spatial overlays. After receiving a standardized explanation of the lumbar puncture procedure, participants began the 20–25 minute training session with either the Proactive AI Facilitator or the Reactive AI Facilitator, based on their assigned group.

Following the training, participants had 15 minutes to perform the same procedure without the headset and without AI support. Afterward, they completed a short post-survey and took part in a 30-minute stimulated recall interview. During the interview, participants viewed short video clips from their session—captured from both the GoPro and Vision Pro—and reflected on key AI interaction moments. They rated each instance on helpfulness, timing, and whether the AI felt supportive or disruptive.



Figure 4: Participant photo in the user study: learning phase (wearing the headset) vs. practical test (not wearing the headset)

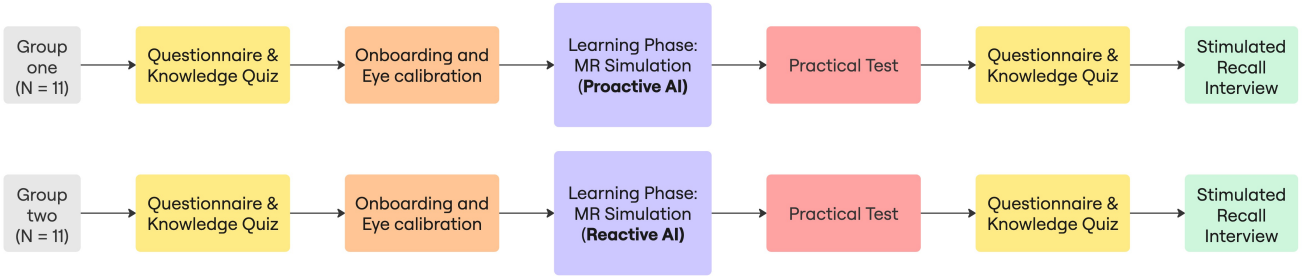


Figure 5: Study procedure flow chart

3.3 Data Collection and Analysis

3.3.1 Quantitative Analysis. To examine the impact of the AI facilitators on learner outcomes in the lumbar puncture training, we collected multiple forms of quantitative data across the study.

Data Collection. At the start of each session, participants completed a pre-training survey designed to establish baseline attitudes and confidence toward AI-facilitated procedural instruction. This survey included an AI Trust scale [17] to measure willingness to rely on AI guidance, a Self-Efficacy scale[85] to capture perceived ability to perform the lumbar puncture, and an expert-developed procedural knowledge quiz that tested foundational understanding of the lumbar puncture procedure, including anatomical landmarks, required equipment, and step order.

After training, participants completed a structured performance assessment in which medical educators used a rubric specifically tailored to the lumbar puncture procedure. This rubric evaluated whether participants completed all of the required steps, such as identifying the correct intervertebral space and inserting the needle correctly; whether the steps were performed in the proper sequence; whether the participant successfully collected cerebrospinal fluid (CSF) into the tube; and whether the CSF was produced using an appropriate technique. This expert-graded assessment provided a structured and objective evaluation of procedural accuracy and skill application beyond self-report.

Following this assessment, participants completed a post-training survey that included the procedural knowledge quiz again, along with the Self-Efficacy scale [85] to evaluate gains in knowledge and confidence. The post-survey also incorporated validated instruments to assess the quality of the learning experience. The User Engagement Scale[69] focused attention subscale measured sustained cognitive involvement during the procedure. The Intrinsic Motivation Inventory subscales [63] for relatedness and pressure/tension captured perceived social connection with the AI facilitator and the level of stress experienced under its guidance. Two workload dimensions from the NASA-TLX[35], temporal demand and frustration, were included to evaluate cognitive load and emotional strain during the lumbar puncture. In addition, participants rated the perceived usefulness of the AI support. For those assigned to the proactive facilitator, the survey also asked whether the timing of its interventions matched their learning pace. These measures created a comprehensive profile of how the AI facilitators influenced skill acquisition, cognitive workload, and learner perceptions within the context of lumbar puncture training.

Data Analysis. We conducted Welch's independent samples t-tests to compare baseline and outcome differences across conditions for all continuous measures, including procedural knowledge gains, self-efficacy change scores, user engagement, workload, interaction counts, and perceived usefulness. We first assessed normality using visual inspection (histograms and Q–Q plots), and no substantial violations were observed, so non-parametric tests were not required. All analyses used two-tailed significance thresholds at $\alpha = .05$. To support interpretation of practical magnitude, we calculated Cohen's d effect sizes for each test. Table 1 reports group means, standard deviations, t-statistics, p-values, and effect sizes for

each comparison between the proactive and reactive AI facilitator conditions.

3.3.2 Qualitative Analysis. To complement the quantitative findings, we conducted an in-depth qualitative analysis that drew on two sources: behavioral video coding and thematic coding of stimulated recall interviews. This approach allowed us to capture both observable patterns in learner behavior and participants' reflective accounts of their experiences with the AI facilitators during lumbar puncture training.

Behavior Coding. Our behavioral coding process was guided by the methodological principles of Interaction Analysis, which emphasize repeated video review, collaborative calibration, and grounding analytic claims in observable evidence [42]. Following these guidelines, the two authors jointly reviewed the first and third training sessions to produce content logs and analytic memos that captured both facilitator–learner conversations and the learner actions that occurred in response. This initial review supported the development of a coding scheme that classified facilitator messages into four categories: welcome or encouragement, check-in guidance, reminder prompts, and learner-initiated questions.

The two authors then independently applied this scheme to the reviewed sessions and met to compare results, refine category definitions, and resolve discrepancies through discussion. Once consensus was established, they independently coded the remaining sessions. For each exchange, the coders documented whether the message was initiated by the facilitator or triggered by a learner request, and they recorded how the learner responded, noting whether the guidance was followed, ignored, or verbally acknowledged. When the interaction was initiated by the learner, their question was further classified as either procedural (e.g., "Do I need to insert the needle any further?") or knowledge-based (e.g., "Will the needle size cause any damage to the ligaments?"). After completing independent coding, the authors met again to consolidate the dataset and ensure consistency across sessions. This systematic process, grounded in the guidelines of Interaction Analysis, allowed us to generate a reproducible account of facilitator–learner interactions that linked AI messages to learner actions within the lumbar puncture training.

Interview coding. Our interview analysis followed Braun and Clarke's thematic analysis framework, which provides a systematic six-phase process for identifying and refining patterns in qualitative data [10]. To begin, the two authors immersed themselves in the dataset by reviewing and transcribing interview recordings, ensuring familiarity with both the depth and breadth of participant accounts. They then jointly coded the first three transcripts to develop and calibrate an initial coding scheme, documenting analytic memos to capture early impressions of learner perceptions of the AI facilitator.

Next, the authors independently applied this coding scheme across the remaining transcripts, generating initial codes that captured recurrent features such as usefulness, timing, tone, initiative, and perceived effects on learning outcomes. In line with Braun and Clarke's guidelines, these codes were iteratively collated into candidate themes, which were then reviewed to ensure internal consistency and clear distinction from one another. Through collaborative refinement, the authors defined and named final themes that captured participants' perspectives at critical junctures of the lumbar puncture procedure, including task initiation and wrap-up [77, 80], high-risk steps such as palpation, anesthetic administration, or needle insertion [29, 87], moments following errors or becoming stuck [29], and periods of waiting, hesitation, or restarting [77]. This process ensured that the analysis was both rigorous and transparent, consistent with Braun and Clarke's recommendation that good thematic analysis grounds interpretive claims in systematic coding while actively situating the researcher's role in identifying patterns.

4 FINDINGS

4.1 Learning Outcomes and Experience

4.1.1 Objective Learning Outcomes. To examine whether a proactive AI facilitator influenced learning outcomes differently than a reactive one, we compared learners' knowledge gains, self-efficacy growth, and post-simulation task performance across two groups. Both groups completed a mixed-reality lumbar puncture training session, but received support from different AI facilitator versions: Group 1 interacted with an AI that offered encouragement, reminders, and help prompts proactively; Group 2 used a reactive version that only responded to direct learner queries.

At baseline, the two groups demonstrated comparable levels of procedural knowledge, self-efficacy, and trust in AI. Pre-training quiz scores did not differ significantly between groups, $t(20) = -0.984$, $p = .337$, $d = -0.41$. AI trust did not show a significant difference, $t(20) = -1.348$, $p = .193$, $d = -0.57$. These results indicate that random assignment produced equivalent starting points across conditions.

Following training, participants in both conditions showed measurable improvement. Quiz scores increased in both groups, and self-efficacy ratings also rose, indicating that the mixed-reality module effectively supported gains in procedural knowledge and confidence. However, the magnitude of these improvements did not differ significantly between the proactive and reactive facilitator conditions. Knowledge gain was statistically comparable across groups, $t(20) = -0.153$, $p = .880$, $d = -0.07$, as was self-efficacy growth, $t(20) = 0.584$, $p = .567$, $d = 0.25$. Despite their differing interaction styles, the two facilitator versions appeared similarly effective in supporting learning gains.

Task performance during the hands-on lumbar puncture simulation likewise showed no significant differences between groups. Participants in both conditions successfully completed most of the seven procedural steps, and performance metrics were statistically comparable across facilitator types. The number of steps completed did not differ significantly, $t(20) = -0.726$, $p = .477$, $d = -0.31$, nor did the correctness of step order, $t(20) = 1.342$, $p = .197$, $d = 0.57$. These results suggest that learners were able to translate instructional content into procedural action effectively, regardless of whether the AI facilitator intervened proactively or only in response to learner queries.

The results suggest that both proactive and reactive AI facilitation led to positive learning outcomes. The level of AI proactiveness did not emerge as a critical factor in determining knowledge gain, confidence building, or procedural success. This suggests that even minimal AI support may be sufficient to foster meaningful learning.

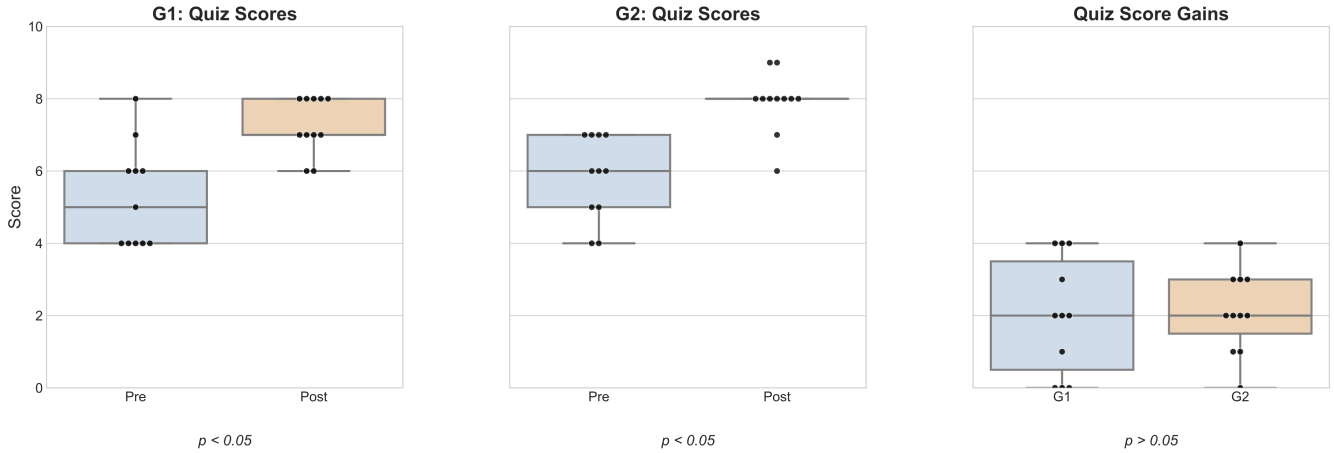


Figure 6: Box plot for knowledge gain of both groups



Figure 7: Box plot for self-efficacy gain of both groups

4.1.2 Subjective Learning Experience. We also assessed learners' subjective experience during the training, including emotional and cognitive aspects of the simulation. Participants rated their engagement, relatedness, pressure/tension, temporal demand, frustration, and perceived usefulness using 5-point Likert scales adapted from validated measures (IMI, NASA-TLX, TAM).

Across all six subjective experience measures, no statistically significant differences were observed between the proactive (Group 1) and reactive (Group 2) facilitator conditions. Engagement scores were similar across groups, $t(20) = 0.457$, $p = .653$, $d = 0.19$. Perceived pressure or tension did not differ meaningfully between conditions, $t(20) = 1.079$, $p = .295$, $d = 0.46$. Workload-related perceptions also showed no differences: temporal demand, $t(20) = -0.343$, $p = .735$, $d = -0.15$. Finally, perceived usefulness of the AI assistance was identical across groups, $t(20) = 0.000$, $p = 1.000$, $d = 0.00$. Full descriptive and inferential statistics are reported in Table 1.

These results indicate that learners experienced both training conditions as similarly useful, engaging, and manageable. Despite the differing facilitator behaviors, participants' emotional and cognitive responses to the training remained consistent across conditions. This suggests that variations in AI proactiveness did not significantly influence learners' overall experience in this immersive training context.

4.2 Behavioral Differences and Interaction Patterns

4.2.1 Quantitative Overview of Interaction Behavior. To explore whether the AI facilitator's proactiveness influenced learner engagement during the simulation, we analyzed the number and types of learner-initiated questions recorded in each group. All interactions with the AI were logged and coded as either procedural, referring to task-specific execution steps, or knowledge-based, focusing on conceptual understanding or clarification.

Table 1: Summary of quantitative comparisons between proactive (G1) and reactive (G2) AI facilitator conditions.

	Group Means and SDs				Cohen's d	t-test	
Comparison	**G1 Mean**	**G1 SD**	**G2 Mean**	**G2 SD**	d	t	p
AI Trust: G1 pre vs G2 pre	2.911	0.626	3.252	0.558	−0.57	-1.348	.193
Self-efficacy: G1 post - pre vs G2 post - pre	2.467	0.939	2.198	1.207	0.25	0.584	.567
Quiz (Normalized): G1 pre vs G2 pre	0.586	0.158	0.646	0.130	−0.41	-0.984	.337
Quiz (Normalized): G1 post - pre vs G2 post - pre	0.222	0.179	0.232	0.126	−0.07	-0.153	.880
User Engagement – Focused Attention	4.394	0.514	4.274	0.711	0.19	0.457	.653
Intrinsic Motivation – Pressure/Tension	2.891	0.288	2.727	0.412	0.46	1.079	.295
NASA-TLX – Temporal Demand	2.273	1.272	2.454	1.214	−0.15	-0.343	.735
Perceived Usefulness	4.091	1.044	4.091	0.831	0.00	0.000	1.000
Step Completed	6.273	0.647	6.455	0.522	−0.31	-0.726	.477
Step Correct Order	5.545	1.128	4.727	1.679	0.57	1.342	.197

Participants in the reactive AI group (Group 2) initiated slightly more interactions overall than those in the proactive group (Group 1). On average, Group 2 learners asked 3.09 questions (SD = 2.21), compared to 1.91 (SD = 1.51) in Group 1; however, this difference was not statistically significant, t(20) = −1.46, p = .158, d = −0.63. A more fine-grained analysis by question type produced similar results. For procedural questions, Group 1 averaged 1.27 (SD = 1.19) while Group 2 averaged 1.91 (SD = 2.12), with no significant difference between groups, t(20) = −0.86, p = .390, d = −0.37. Knowledge-based questions followed the same pattern: Group 1 asked an average of 0.73 (SD = 1.19), compared to 1.18 (SD = 1.47) in Group 2, again yielding no significant difference, t(20) = −0.81, p = .420, d = −0.34. Together, these results indicate that facilitator proactiveness did not meaningfully influence learners' propensity to seek help or clarification during the simulation.

These findings suggest that the proactive AI facilitator, despite providing more frequent unsolicited guidance, did not result in more learner-initiated engagement. If participants in the reactive condition asked more questions, it may be because they were more accustomed to depending on the AI as a responsive tool when uncertainty arose. Still, none of the observed differences reached statistical significance, echoing the pattern found in Section 4.1: high-level engagement metrics remained stable across both versions of the AI facilitator. However, as the next section will show, the surface-level similarity in frequency masks bigger structural differences.

4.2.2 Learner Engagement by Key Moment. To further examine learner engagement patterns beyond the quantitative interaction frequency, we analyzed the timing and context of each learner-initiated interaction with the AI facilitator. Specifically, we coded every spoken or typed input from participants by identifying the key moment within the simulation when the interaction occurred. Each instance of learner speech directed toward the AI was timestamped and then inductively coded based on its temporal and task relevance within the procedure. This led to the emergence of four recurring "key moments" that spanned the training session: (1) Step Start / Step Completion, (2) High-Stakes Steps, (3) Pause / Waiting Periods, and (4) After a Failed Attempt.

Each key moment encapsulated a different type of learner need and intention. For instance, questions at Step Start / Completion often reflected a desire for confirmation or reassurance before progressing. In contrast, questions during High-Stakes Steps, such as needle insertion or anesthetic delivery, were typically aimed at clarifying spatial accuracy or procedural rationale. Pause / Waiting Periods were more exploratory or reflective in nature, as learners used the downtime to revisit prior steps or inquire about procedural details. Finally, interactions that occurred after a Failed Attempt showed learners actively seeking to recover or troubleshoot, often in response to perceived error or AI feedback.

Across all 22 participants, we identified a total of 43 learner-initiated interactions, distributed fairly evenly between the two groups: 22 in Group 1 (proactive AI) and 21 in Group 2 (reactive AI). This finding mirrors the quantitative results reported earlier, where no significant difference in interaction frequency was observed between conditions. Importantly, the distribution of key moments across groups was also comparable. Both groups most frequently initiated interactions during High-Stakes Steps, followed by Pause / Waiting Periods. This suggests that learners across conditions tend to seek help in more complex steps and moments of uncertainty.

Beyond frequency, the content and tone of learner-initiated interactions varied across key moments. During step transitions, questions like "Do I need to insert the needle further?" (P1) or "Should I re-insert the middle part of the spinal needle?" (P22) revealed learners' need for procedural confirmation. In high-stakes moments, questions shifted toward conceptual clarity and confidence, such as P19 asking, "What does aspirating mean?" or P23 asking, "Should I get the needle right now?" to seek reassurance before action. During troubleshooting, learners voiced embodied uncertainty, as P10 noted, "I think I cannot feel the pop," while P2 asked, "Where should I position my hand?" Finally, after failed attempts, learners continued to reflect and check for safety, as in P2's "Will the needle size cause any damage?" These utterances highlight the AI's role as a just-in-time cognitive and emotional scaffold across varying training phases.

While the key moment coding reveals when and why learners sought help, it does not fully capture how these interactions

Key Moment	Dimension	G1	G2	Procedure	Knowledge	Example Conversation
Step Start / Step Completion	Pre-Action Caution	8	7	18	14	P1: "Do I need to insert the needle further?"
	Step-by-Step Guidance					P22: "Should I re-insert the middle part of the spinal needle in between each tube collection?"
	Tool and Equipment Support					P12: "Which needle is used to draw the lidocaine and which one is used to inject the lidocaine?"
	Visual-Spatial Anchoring					P3: "How do you tell where L3 L4 or L4 L5 interace is?"
High-Stakes Step	Concept Reassurance	5	5	5	12	P19: "What does aspirting mean?"
	Tool Refresher					P15: "Which one is the 25 gauge needle?"
	Visual Landmark Reminder					P9: "Can you show me the imaginary line again?"
	Procedural Confidence Prompt					P23: "Should I get the needle right now?"
Pause/Waiting Period	Troubleshooting Tactile Feedback	1	2	2	2	P10: "I think I cannot feel the pop when I inserted deep."
	Repositioning Guidance					P2: "I'm not feeling any changes to resistance um where should I position my hand?"
After a Failed Attempt	Post-Step Confirmation	0	2	2	2	P10: "So that means that I just put a towel like under and besides the patient to catch the fluids."
	End-of-Step Conceptual Safety Check					P2: "Will the needle size cause any damage to the ligaments?"

Figure 8: User-initiated interaction through key moment dimension

unfolded over time or how they shaped the learner's broader experience. To complement this temporal gap, we created user journey maps that visualize the full session trajectories that highlight interaction density, directionality, and learner behaviors. These maps allow us to explore variation within and across groups and surface patterns not evident in aggregate counts.

4.2.3 Learner-AI Interaction Journeys. While statistical analyses revealed minimal differences in interaction frequency between proactive and reactive AI facilitator conditions, the structure and trajectory of those interactions varied substantially from learner to learner.

To more deeply examine how AI proactivity shaped learners' unfolding experience, we conducted a qualitative trajectory analysis by generating "learner–AI interaction journey maps" for all 22 participants. These maps visualize temporal patterns in (1) who initiated each exchange (AI or learner), (2) the procedural moment in which it occurred (e.g., Greeting, Encouragement, Check-in Guidance, Reminder Question, Learner Question), and (3) the learner's post-study perception towards a certain interaction, where we also annotated their sentiment markers (green = positive, yellow = neutral, red = negative). By examining who initiated each exchange, how participants responded, and how their behavior evolved over time, we surfaced several critical differences in learner engagement that were not captured in survey metrics.

From these 22 maps, we selected four participants: P22 and P8 from the Proactive AI condition, and P2 and P10 from the Reactive AI condition, to illustrate the qualitatively distinct interaction patterns that appeared across the dataset. Crucially, these participants were not chosen because their data were more positive or more dramatic, but because their trajectories each represent a different, recurrent pattern we observed across multiple learners: High-frequency, reflective help-seeking, represented by P22 (Fig. 9, left) and also seen in 4 other participants; Minimal-interaction, task-focused progression, represented by P8 (Fig. 9, right) and seen in 5 other participants; Iterative, post-completion clarification, represented by P2 (Fig. 10, left) and observed in 3 other participants; and Linear progression with occasional, step-specific clarification, represented by P10 (Fig. 10, right) and observed in 6 other participants).

Learners rarely responded to AI-initiated interactions. One consistent pattern was the very low rate of verbal responses to AI-initiated interventions. Although the proactive AI was designed to offer check-ins, reminders, and encouragement at carefully selected moments, most learners treated these messages as background guidance rather than conversational turns. As illustrated in Fig. 9, only one proactive prompt, Block 5 in P22's session, elicited an explicit verbal response when the learner paused and answered the AI's question. In contrast, P8 ignored all three AI-initiated prompts, including two reminders and a check-in at blocks 4, 7, and 10, continued the task uninterrupted.

In the reactive condition (G2), such proactive prompts were absent by design, but the interaction blocks consistently featured one-turn exchanges, even when learners initiated the questions. None of the participants followed up after the AI's response, suggesting that single-turn interactions may be a general behavioral pattern in this setting, regardless of who initiates the conversation.

These patterns contextualize our quantitative finding that interaction frequency did not differ significantly across conditions: although the proactive AI spoke more often, learners did not treat its messages as openings for dialogue. Instead, both groups tended to use the AI in momentary, task-specific transactions, suggesting that conversation-like turn-taking may not be the dominant form of interaction in embodied procedural training.

Learners initiated interaction primarily for procedural clarification. While overall rates of learner-initiated questions were similar across conditions, the journey maps reveal distinct motivational patterns, shaped by learners' disposition and task approach.

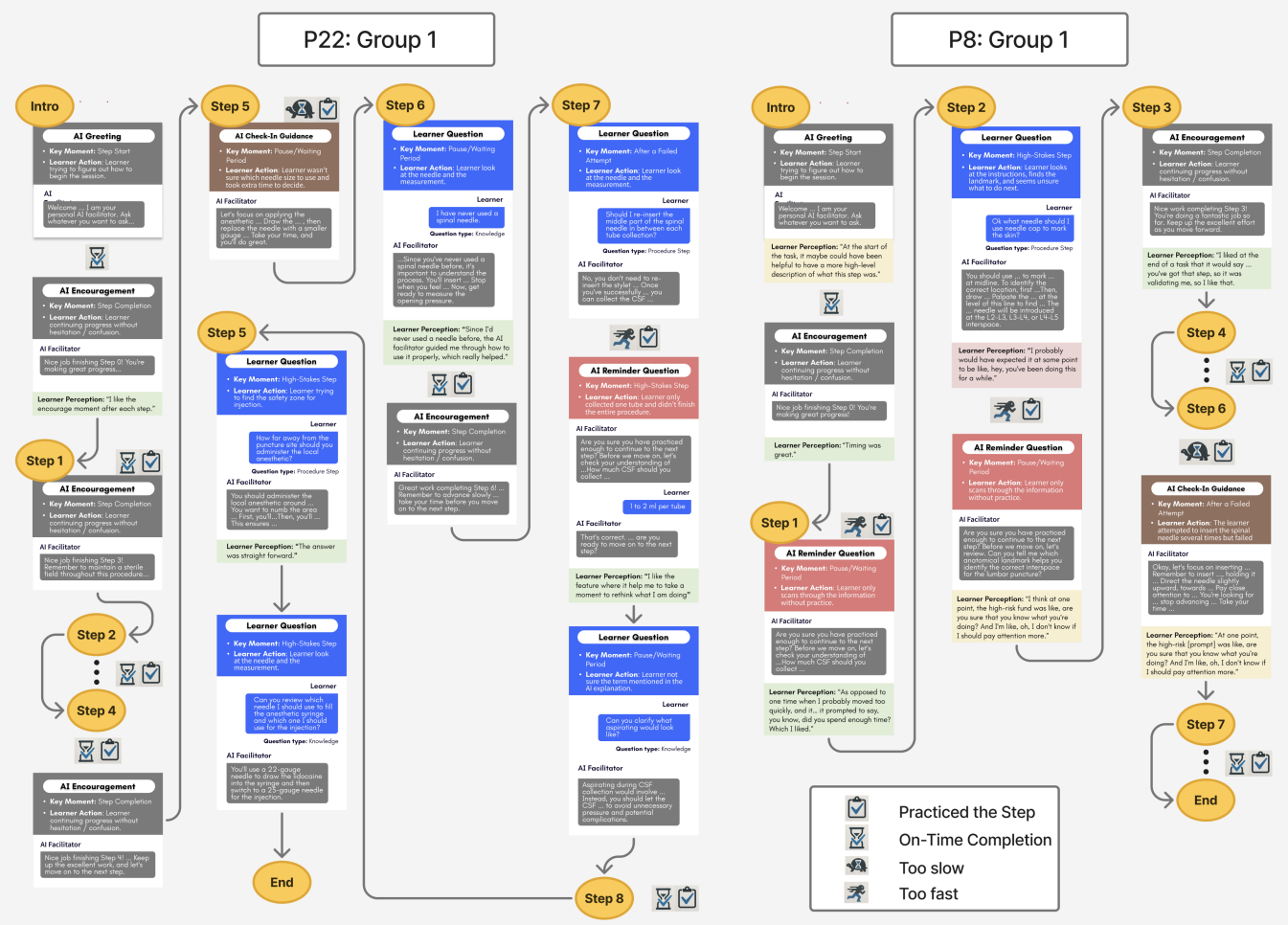


Figure 9: User journey map for P22 and P8 from group 1

In the proactive condition (Fig. 9), P22 exemplifies a highly reflective pattern – initiating seven help-seeking interactions across the tutorial. This learner initiated seven questions covering both procedural and conceptual clarifications, especially during transitions or decision points in the task (e.g., *"How far away from the puncture site should you administer the anesthetic?" and "Can you clarify what aspirating would look like?"*). After completing the procedure, this learner returned to earlier steps to ask additional clarifying questions, suggesting a more exploratory and reflective approach to the tutorial. While the AI was proactive in this condition, many of these learner-initiated questions occurred independently of system prompts. By contrast, P8 reflects a focused, task-driven pattern observed in several other learners. P8 initiated only one question, at a moment of hesitation, and otherwise maintained steady forward progress without revisiting earlier steps.

Similar contrasts appeared in the reactive condition (Fig. 10). P2 also initiated seven questions, including both procedural and conceptual questions. Notably, after completing the full procedure, P2 returned to earlier steps to ask follow-up questions, as seen in the final blocks of their journey map. While each question was a single-turn exchange, the sequence reflects an intent to consolidate knowledge and revisit moments of uncertainty. In contrast, P10 showed a goal-oriented but clarification-seeking behavior. This type of learner moved linearly from start to finish without revisiting earlier steps, but still initiated three procedural questions, two of which involved multi-turn exchanges.

Across both conditions, these representative cases demonstrate that learner disposition, whether reflective, exploratory, or task-oriented, played a more decisive role in help-seeking behavior than whether the AI was proactive or reactive.

Timeline patterns and interaction density reveal distinct trajectories. The journey maps further highlight common trajectory types found across participants: Patterns varied in whether learners progressed linearly or revisited prior steps, and whether AI interventions were concentrated at specific points or spread more evenly across the session. The selected participants exemplify these patterns, but each pattern appeared in multiple sessions.

For example, in the proactive condition (Fig. 9), P22 followed a looping trajectory. Although the learner completed all eight steps, the timeline showed detours shaped by AI proactivity. P22 stayed

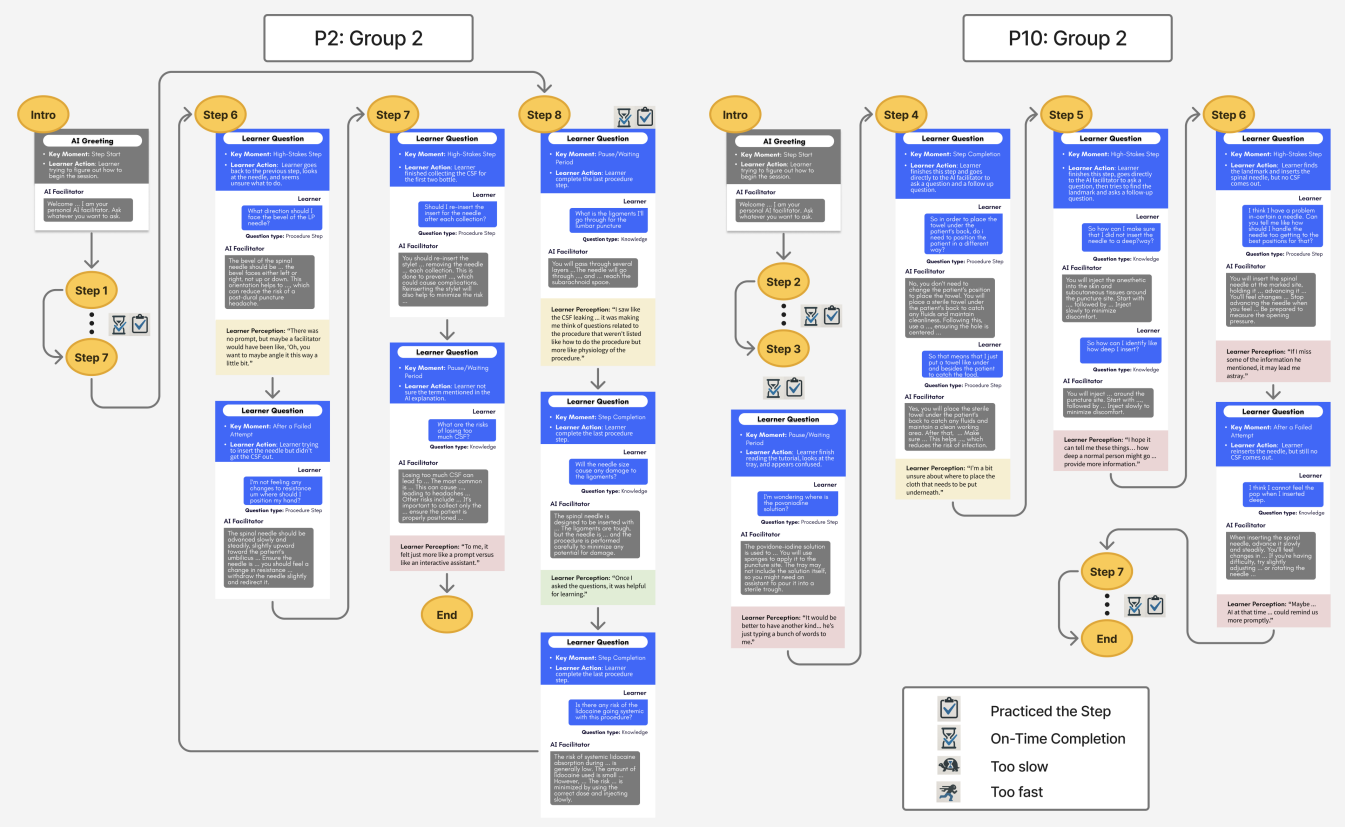


Figure 10: User journey map for P2 and P10 from group 2

at Step 5 for an extended time and triggered an AI check-in. And at Step 7, P22 moved too quickly and activated an AI proactive reminder question. After finishing Step 8, the learner returned to Step 5 and asked two additional procedural questions, then revisited Step 7 for further clarification. The result was a dense timeline with multiple points of re-engagement. P8, by contrast, displayed a linear trajectory with two early corrective AI interventions. At Step 1 and Step 2, moving too quickly triggered two proactive AI reminders. After these early prompts, the learner slowed down and advanced through the remaining steps without further speed-related triggers. This suggests that AI feedback could shape subsequent pacing.

In the reactive condition (Fig. 10), P2 also circled back after completion with AI intervention, but in a different manner from P22. The learner progressed through the steps with minimal AI involvement, then returned post-completion to Step 6, asking two questions, and to Step 7, asking two more. This created an episodic density curve, with the majority of interactions clustered after the nominal end of the tutorial. In contrast, P10 contrasted with these looping patterns by remaining strictly linear. P10 interacted with the AI facilitator at several steps, but without backtracking. This produces a steady progression through the tutorial with a moderate distribution of questions across Steps 2–6. These trajectories highlight how AI proactivity and learner pacing produced distinct experiential rhythms. Group 1 learners (P22, P8) encountered branching timelines shaped by proactive AI checks and reminders, while Group 2 learners (P2, P10) demonstrated more self-directed paths.

Across these representative trajectories, we see how AI proactivity, pacing, and learner strategy jointly shaped the experiential rhythm of the tutorial. Proactive prompts influenced pacing for some learners (e.g., P8), whereas others (e.g., P22, P2) exhibited reflective revisiting independent of AI initiative. This diversity highlights the value of analyzing trajectories rather than relying solely on aggregate statistics.

4.3 Learner Perceptions of AI Facilitation

While the prior section analyzed learner-AI interaction patterns, we further explored how learners in both groups perceived the AI facilitator's role through post-session interviews. Their reflections surfaced recurring themes related to who should initiate interactions, how well-timed the interventions were, and whether the moments chosen by the AI felt appropriate or necessary.

4.3.1 Control Over Interaction Flow. A central theme that emerged across both groups was the question of who should lead the interaction—the learner or the AI facilitator. While Group 1 participants experienced a proactive AI that sometimes initiated prompts, and Group 2 participants only engaged the AI reactively by choice, learners in both conditions expressed preferences that did not always align with the system they used.

In the proactive condition, several learners noted discomfort when the AI spoke without being prompted. One participant reflected, *"It was a little too much. A little annoying when it keeps on talking, and you're like, I already know I want to go to the next step."* (P6). Another shared a similar sentiment: *"I think it's annoying. I think… if you're thinking about the AI, it's like… I see, so you should just be on the side, and just don't bother."* (P7). These responses suggest that even when support is available, learners may value the freedom to decide when and how to engage. However, this preference for learner-led control was not universal. Some participants in Group 1 appreciated the AI's proactive behavior and described it as helpful and encouraging. One participant explained, *"Yeah, I definitely think the proactive did encourage me to actually use the AI"* (P1). They further elaborated on the benefit of the AI's guidance: *"I think it helped to refocus. I don't think I felt rushed. I think I… like, I… it definitely, like, pushed me to use it, but I think that I needed to be pushed to use it, because I was very, like, apprehensive in the beginning."* (P5). For this learner, the AI's initiative helped scaffold the learning process without undermining their autonomy.

In Group 2, which featured a reactive AI that responded only when prompted, some learners reflected on moments when they hesitated to ask for help and wished the system had offered unsolicited guidance. One participant noted, *"Like, after palpating the spine to find where to put the needle, maybe the AI could prompt me: 'Okay, now try to insert the needle.' And then it could say if the needle was in the incorrect location or depth."* (P12). Another added, *"I do want them at the high-risk step at five, and then right here after an error or when stuck. This is, of course, important for the AI as a teaching tool."* (P18). These reflections point to a potential mismatch between learner needs and the burden of initiating interaction, particularly in situations where uncertainty or hesitation might prevent learners from reaching out. The contrasting responses illuminate the importance of designing for varying degrees of learner control. Some learners thrive with an AI that takes initiative. Others prefer to retain full control over when and how assistance is delivered. Notably, these preferences did not always align with the AI behavior they encountered in the study, where participants expressed desires to have features displayed in the other condition.

4.3.2 Experience of Temporal Fit. While the AI facilitator was designed to offer timely support, learners frequently remarked on the pacing of its interventions. The timing of the AI, whether too early, too late, or just right, shaped participants' overall experience and their ability to stay engaged in the task. Some learners appreciated the AI's timing when it aligned with their needs. One participant noted, *"As opposed to one time when I probably moved too quickly, and it… it prompted to say, you know, did you spend enough time? Which I liked."* (P8). Another added, *"I recall that when inserting, I got stuck a few times at the beginning… when it guided me, it provided logical steps, like saying: 'Stop advancing when you feel a pop, that means you're in the subarachnoid space.' That was useful.* (P10). This suggests that timely nudges were effective in re-engaging attention.

However, not all interactions were perceived as well-timed. Several learners described moments where the AI interrupted their thought process. One participant explained, *"The timing sometimes felt disruptive. For example, when I was angling the needle, it told me I was taking a long time. That distracted me."* (P2). Another reflected, *"…there are a variety of reasons as to why there is a waiting or pause period. It could be that the user is just organizing their thoughts during this time. It doesn't necessarily mean that they're stuck in the step. And so AI might be seen as a distraction if it initiates contact with the user at that time."* (P18). The perceived misalignment between the AI's pacing and the learner's internal rhythm led to cognitive friction. Overall, learners are highly sensitive to timing, both in terms of AI-initiated and learner-initiated interactions. While proactive prompts can be helpful when well-timed, even slight misalignment with a learner's mental rhythm or task engagement can lead to frustration.

4.3.3 Expectations About Situational Relevance. Learners frequently reported the AI's helpfulness based on whether the moment felt situationally appropriate. Their perceptions were shaped by the complexity, emotional weight, or criticality of a given step. In steps perceived as confusing or high-stakes, learners welcomed guidance. But during simpler moments, the same intervention could feel unnecessary or distracting.

Participants often described heightened expectations for AI support during difficult moments. One learner remarked, *"Like when I was getting stuck on something, like on a certain step, I can ask it a question"* (P1). This suggests that cognitive demand increased the perceived value of assistance. Another noted, *"So it really did force me to ask questions rather than just me thinking to myself... I was like, you know, having a mentor in the room"* (P3). This indicates that the AI's perceived role shifted based on task complexity.

In contrast, learners expressed discomfort when the AI engaged in low-risk or straightforward moments. One participant said, *"More helpful at the end of harder steps, but if it's just putting on the drape or gloves, then it should be quick"* (P15). Another echoed this sentiment: *"It was really easy and straightforward to follow. I just did it without thinking too much"* (P9). Even when the AI's timing was not problematic, learners questioned its judgment when it intervened during tasks that didn't warrant support. As one put it, *"I think there's a difference between when you're confused versus when you're just going through the motions"* (P9). These reflections underscore the need for AI facilitators to interpret not just procedural steps, but the situational relevance of those steps in relation to learner needs. The same action might be routine for one learner and confusing for another.

5 DISCUSSION

5.1 Proactive AI alone did not affect the learning outcome in immersive medical training

As reported in Section 4.1, we found no statistically significant differences in learning outcomes between participants who trained with a proactive AI facilitator and those who used a reactive one. This finding parallels prior work showing that unsolicited AI assistance can sometimes reduce learner satisfaction or interfere with autonomy. Diebel et al. found that proactive help could lower user satisfaction if it disrupted task flow and made participants feel observed or second-guessed [20]. In the context of immersive medical training, where learners are immersed in embodied activity, this disruption may be especially critical. Learners in our study described

some proactive interventions as helpful, but others perceived them as unnecessary or mistimed, particularly during lower-intensity procedural phases.

Similar effects have been observed in educational contexts where AI monitors learner behavior and intervenes based on inferred struggle. Kraus et al. reported that proactive interventions based on affective state sometimes caused learners to feel rushed or under pressure [48]. This suggests that the value of proactivity depends on the learner's readiness to receive help. Learner identity also played an important role. In professional training contexts, unsolicited guidance can challenge learners' perceptions of competence. Goutier et al. found that proactive support increased self-threat among users completing identity-relevant tasks, such as medical decision making [31]. Our participants, most of whom were early-career medical students, may have interpreted proactive interventions as a signal of underperformance.

Finally, the design of our system itself may have moderated the effects of facilitator initiative. Both conditions included spatial overlays, step-specific guidance, and real-time feedback presented within the learner's visual field. These XR features likely provided a strong baseline of support. The proactive facilitator relied on behavioral thresholds to trigger feedback, but this did not always align with the learner's internal needs or timing preferences. These findings suggest that the effectiveness of proactive facilitation is not guaranteed.

5.2 Design Boundaries of Proactive Facilitation

The findings presented in Sections 4.2 and 4.3 show that the effects of proactivity were not uniform. Learners did not simply prefer one version of the AI facilitator over the other; rather, they evaluated each intervention through expectations of agency, timing, and situational relevance. These evaluations reveal recurring tensions between moments when proactivity was motivating and moments when it was perceived as intrusive or unnecessary. To provide actionable guidance for designing AI facilitators in mixed-reality (MR) training systems, we distill these tensions into three boundaries: the Initiative Boundary, the Temporal Boundary, and the Situational Boundary. Together, they articulate how proactive behaviors should be calibrated to support learners within embodied, hands-on workflows.

5.2.1 Initiative Boundary: Who Leads the Interaction? The first boundary concerns the degree to which the AI should initiate interactions rather than waiting for learner input. As discussed in Section 4.3.1, learners held divergent expectations about who should lead the flow of interaction. Our results suggest that initiative was not experienced as a simple binary between proactive and reactive facilitation, but as a continuum requiring careful calibration. Some learners value AI initiative as a way to scaffold participation and reduce hesitation, while others prefer to maintain full control over when to seek assistance. For MR-based training systems, this boundary highlights a practical design implication: initiative should not be fixed at design time [9, 40]. Instead, systems may benefit from: 1) learner-adjustable initiative settings, similar to mixed-initiative control in tutoring systems; or 2) adaptive initiative models that infer when learners prefer autonomy (e.g., fluent performance, confident pacing) versus support (e.g., extended hesitation or uncertainty). The Initiative Boundary foregrounds how MR systems must balance structured guidance with a learner's desire for agency, a balance that becomes especially delicate in embodied environments where users are physically engaged and cannot rely solely on verbal cues to negotiate control.

5.2.2 Temporal Boundary: When Is the Right Moment to Intervene? The second boundary concerns the pacing and synchronicity of AI interventions. As seen in Section 4.3.2, learners were highly sensitive to whether the AI's responses felt timely. Interventions were appreciated when they coincided with hesitation or genuine confusion, but they were disruptive when delivered prematurely. Behavioral analysis showed that proactive reminders occasionally led some learners to adjust their pacing, whereas others disengaged from the AI altogether when the timing clashed with their workflow. For MR-based systems, this boundary underscores a core implication: timing matters even more in immersive environments, where learners are coordinating hand movements, gaze, and spatial navigation. Designing for this boundary may require: 1) pacing-sensitive models that adjust interventions based on variations in movement speed, hesitation, or repeated actions; 2) temporal buffers to prevent the AI from interrupting ongoing motor sequences [19]; and 3) multi-signal temporal inference combining gaze, hand trajectories, and tool manipulation to detect appropriate intervention windows [75]. The Temporal Boundary emphasizes that effective proactivity is not a matter of frequency but of alignment with the embodied tempo of the learner's activity.

5.2.3 Situational Boundary: When Is Proactivity Appropriate? The third boundary addresses the situational relevance of AI interventions. As discussed in Section 4.3.3, learners also evaluated interventions on whether they were appropriate for the type of task at hand. Participants welcomed proactive input during complex or high-stakes steps. Conversely, interventions during routine or straightforward actions were dismissed as unnecessary. This distinction was reinforced by behavioral analysis of key moments (Section 4.2). Learners frequently expected support after failed attempts or during high-stakes steps, but often preferred silence during routine transitions or while reflecting in pauses. These patterns suggest that appropriateness depends on the perceived criticality and cognitive demand of the moment. The Situational Boundary emphasizes the importance of calibrating proactivity to the task context. Designers should distinguish between moments where intervention is expected. Context-sensitive facilitation may require systems that incorporate task tagging, learner state estimation, or adaptive heuristics for gauging complexity.

These three boundaries represent distinct dimensions along which AI behavior must be calibrated. Importantly, these dimensions intersect with the four key moments identified in behavioral analysis (Section 4.2): the start or end of a step, high-stakes steps, pauses or waiting periods, and the aftermath of an error. Learners evaluated the AI differently across these moments. To illustrate how these findings can inform future design, we synthesize them into a design table (Table 2). Each row corresponds to a recurring moment in the procedural workflow, and each column articulates how the three boundaries should be tuned for that moment. The table, therefore, serves as a practical reference that designers can

Key Moment	Initiative	Temporal	Situational
Step Start / End	Some learners welcomed proactive orientation, others preferred to initiate themselves.	Avoid overlapping with instructions; wait until learners have consumed the step material.	Provide lightweight confirmation or navigation support rather than heavy guidance.
High-Stakes Step	Proactive initiative is often expected, as learners may hesitate to request help.	Timely feedback during execution is critical; avoid premature prompts that disrupt focus.	Interventions here are perceived as highly valuable and appropriate.
Pause / Waiting	Learners often prefer to lead; proactive initiative during silence may feel intrusive.	Avoid immediate responses; allow hesitation to unfold before nudging.	Relevance is low unless hesitation is prolonged, in which case a gentle check-in is appropriate.
After Error	Learners strongly expect the AI to step in unprompted for correction.	Quick but not repeated feedback is valued; excessive repetition causes frustration.	Interventions are seen as highly appropriate, reinforcing correction and recovery.

Table 2: Design table synthesizing behavioral analysis of key moments (Section 4.2) with learner perceptions of AI proactivity (Section 4.3). Each row represents a training moment, and each column shows how initiative, timing, and situational relevance should be calibrated for that moment.

consult when deciding how and when an MR-based AI facilitator should intervene.

For example, after an error, the table indicates that learners strongly expect the AI to initiate support, and that such intervention is perceived as highly appropriate. An MR training system could therefore be designed to automatically trigger a brief corrective prompt, such as highlighting the correct needle orientation in 3D space, immediately after the learner performs an incorrect action. In contrast, during pauses or waiting, the Initiative and Situational Boundaries suggest restraint: the AI should hold back unless hesitation becomes prolonged, at which point a gentle check-in is more suitable than a directive correction. This distinction helps prevent unnecessary interruptions while still offering support when learners need it. By mapping boundaries to moments, the table offers a concrete, moment-by-moment framework for calibrating AI facilitation in immersive training systems, enabling designers to implement proactive behaviors that are both pedagogically supportive and aligned with learners' expectations in MR environments.

5.3 Limitations and Future Work

While our study provides new insights into the role of proactivity in AI facilitation, several limitations should be acknowledged. First, the study was conducted in the context of a single procedural skill within a simulation-based training environment. Although the task represents a high-stakes and complex procedure, future research should examine whether the boundaries we identify generalize to other domains. Second, the proposed proactive triggers timing threshold was developmental, serving as design probes to explore learner perceptions of proactive facilitation rather than as validated indicators of struggle or readiness. Readers should note that different threshold designs may yield different facilitation dynamics. Future work could refine these heuristics through adaptive or data-driven calibration. Third, we recruited 22 participants, which aligns with sample sizes commonly used in mixed-methods evaluations of immersive AI systems [52, 54]. This sample size balances statistical power with the practical constraints of medical student availability and the time-intensive XR laboratory setup. The study is powered to detect medium-to-large between-subject effects, and was not powered to detect small quantitative differences. Fourth, while our findings integrate both behavioral data and self-reported perceptions, the study was limited to a single session per learner. Prolonged exposure to proactive facilitation may alter perceptions over time, as learners develop trust or adapt to system behaviors. Longitudinal studies are suggested to understand how boundaries of initiative, timing, and situational appropriateness shift with repeated training. Finally, our participant sample consisted of medical students in one institutional setting. Broader evaluation across institutions, levels of expertise, and cultural contexts will be necessary to validate the generalizability of our framework. Despite these limitations, our findings contribute to a design space that bridges behavioral dynamics and learner perceptions. Future research can extend this space by exploring adaptive mechanisms that modulate proactivity in real time.

6 CONCLUSION

This study examines how the proactiveness of AI facilitators shapes learner experience in immersive medical training. By comparing a proactive and a reactive version of an AI-enabled lumbar puncture simulator, we showed that differences in proactivity did not yield measurable outcome differences but produced distinct patterns of interaction and divergent learner perceptions. Through behavioral trajectory analysis and post-study interviews, we identified recurring moments when interventions clustered and revealed tensions in how learners judged initiative, timing, and appropriateness.

From these findings, we proposed three boundaries, the Initiative, Temporal, and Situational Boundaries, that define how proactivity

should be calibrated in AI facilitation. Crossing these boundaries with four key moments of training derived from behavioral analysis, we synthesized a design table that provides practical guidance for when and how AI facilitators should engage. This framework highlights that proactivity is a multidimensional design problem that should balance learner autonomy, task rhythm, and situational relevance.

While our study was situated in the context of procedural skill training for lumbar puncture, the implications extend to other domains where AI is introduced as a facilitator in immersive environments. By integrating quantitative, behavioral, and qualitative data, this work contributes a grounded framework for designing AI facilitators that are context-sensitive, learner-aware, and adaptable across training contexts. We hope these insights inform future research on adaptive facilitation and spark new explorations of how AI systems can support learning with consistent, contextually attuned guidance.

REFERENCES

[1] Shashank Ahire, Benjamin Simon, and Michael Rohs. 2024. WorkFit: designing proactive voice assistance for the health and well-being of knowledge workers. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. 1–14.

[2] Stephen Aiello, Thomas Cochrane, and Charles Sevigny. 2023. The affordances of clinical simulation immersive technology within healthcare education: a scoping review. *Virtual Reality* 27, 4 (2023), 3485–3503.

[3] Abdulmohsen H Al-Elq. 2010. Simulation-based medical teaching and learning. *Journal of family and Community Medicine* 17, 1 (2010), 35–40.

[4] Guillaume Alinier and Denis Oriot. 2022. Simulation-based education: deceiving learners with good intent. *Advances in Simulation* 7, 1 (2022), 8.

[5] Tatiana Bachkirova and Rob Kemp. 2025. 'AI coaching': democratising coaching service or offering an ersatz? *Coaching: An International Journal of Theory, Research and Practice* 18, 1 (2025), 27–45.

[6] Saptarashmi Bandyopadhyay, Vikas Bahirwani, Lavisha Aggarwal, Bhanu Guda, Lin Li, and Andrea Colaco. 2025. YETI (YET to Intervene) Proactive Interventions by Multimodal AI Agents in Augmented Reality Tasks. *arXiv preprint arXiv:2501.09355* (2025).

[7] Anil Bhattad, Monali Gulhane, Kiran Gavhale, Abhay Revatkar, Dilip Patil, and Uday Pratap Singh. 2024. Generative AI for Simulating Rare Disease Scenarios in Training Robots. In *Proceedings of the 6th International Conference on Information Management & Machine Intelligence*. 1–10.

[8] Dan Bohus, Sean Andrist, Nick Saw, Ann Paradiso, Ishani Chakraborty, and Mahdi Rad. 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)*. IEEE, 889–890.

[9] Jeffrey M Bradshaw, Paul J Feltovich, Hyuckchul Jung, Shriniwas Kulkarni, William Taysom, and Andrzej Uszok. 2003. Dimensions of adjustable autonomy and mixed-initiative interaction. In *International workshop on computational autonomy*. Springer, 17–39.

[10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[11] Claire L Burns. 2015. Using debriefing and feedback in simulation to improve participant performance: an educator's perspective. *International journal of medical education* 6 (2015), 118.

[12] Mukundan Chariar, Shreyas Rao, Aryan Irani, Shilpa Suresh, and CS Asha. 2023. AI trainer: Autoencoder based approach for squat analysis and correction. *IEEE Access* 11 (2023), 107135–107149.

[13] Valerie Chen, Alan Zhu, Sebastian Zhao, Hussein Mozannar, David Sontag, and Ameet Talwalkar. 2025. Need help? designing proactive ai assistants for programming. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[14] Xinyue Chen, Lev Tankelevitch, Rishi Vanukuru, Ava Elizabeth Scott, Payod Panda, and Sean Rintel. 2025. Are We On Track? AI-Assisted Active and Passive Goal Reflection During Meetings. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.

[15] Yang Chen, Katherine Fennedy, Jiayi Zhang, Yong Jie Sim, Clement Zheng, and Ching Chiuan Yen. 2025. Bridging Simulation and Reality: Augmented Virtuality for Mass Casualty Triage Training-From Landscape Analysis to Empirical Insights. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–28.

[16] Diana-Margarita Córdova-Esparza. 2025. Ai-powered educational agents: Opportunities, innovations, and ethical challenges. *Information* 16, 6 (2025), 469.

[17] Edoardo Sebastiano De Duro, Giuseppe Alessandro Veltri, Hudson Golino, and Massimo Stella. 2025. Measuring and identifying factors of individuals' trust in Large Language Models. *arXiv preprint arXiv:2502.21028* (2025).

[18] Yang Deng, Lizi Liao, Wenqiang Lei, Grace Hui Yang, Wai Lam, and Tat-Seng Chua. 2025. Proactive conversational ai: A comprehensive survey of advancements and opportunities. *ACM Transactions on Information Systems* 43, 3 (2025), 1–45.

[19] Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. 2024. Towards human-centered proactive conversational agents. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 807–818.

[20] Christopher Diebel, Marc Goutier, Martin Adam, and Alexander Benlian. 2025. When AI-Based Agents Are Proactive: Implications for Competence and System Satisfaction in Human–AI Collaboration. *Business & Information Systems Engineering* (2025), 1–20.

[21] Dinithi Dissanayake and Suranga Nanayakkara. 2025. Navigating the State of Cognitive Flow: Context-Aware AI Interventions for Effective Reasoning Support. *arXiv preprint arXiv:2504.16021* (2025).

[22] Lauren Dugan, Kyle Rei, Arisa Ueno, Joshua Mahutga, Timothy Hantz, and Madhu Varma. 2023. Assessing Best Practices in a Simulated Lumbar Puncture Workshop with Medical Students (P4-2.004). *Neurology* 100, 17_supplement_2 (2023), 1475.

[23] Vanessa Echeverria, Linxuan Zhao, Riordan Alfredo, Mikaela E Milesi, Yueqiao Jin, Sophie Abel, Jie Xiang Fan, Lixiang Yan, Samantha Dix, Rosie Wotherspoon, et al. 2025. TeamVision: An AI-powered Learning Analytics System for Supporting Reflection in Team-based Healthcare Simulation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.

[24] Chukwuka Elendu, Dependable C Amaechi, Alexander U Okatta, Emmanuel C Amaechi, Tochi C Elendu, Chiamaka P Ezeh, and Ijeoma D Elendu. 2024. The impact of simulation-based training in medical education: A review. *Medicine* 103, 27 (2024), e38813.

[25] Renaud Felten, Kevin Bigaut, Thomas Wirth, Laurent Kremer, Lucas Gauer, Cécile Arnold, Irène Ollivier, Julien Godet, Marc Scherlinger, Maxime Dubois, et al. 2025. Advancing medical training with augmented reality and haptic feedback simulator: outcomes of a randomized controlled trial on lumbar puncture. *BMC Medical Education* 25, 1 (2025), 1231.

[26] Kristin L Fraser, Michael J Meguerdichian, Jolene T Haws, Vincent J Grant, Komal Bajaj, and Adam Cheng. 2018. Cognitive load theory for debriefing simulations: implications for faculty development. *Advances in Simulation* 3, 1 (2018), 28.

[27] Mingyu Fu, Wei Fang, Shan Gao, Jianhao Hong, and Yizhou Chen. 2022. Edge computing-driven scene-aware intelligent augmented reality assembly. *The International Journal of Advanced Manufacturing Technology* 119, 11 (2022), 7369–7381.

[28] Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–24.

[29] David Gent and Ranjev Kainth. 2022. Simulation-based procedure training (SBPT) in rarely performed procedures: a blueprint for theory-informed design considerations. *Advances in Simulation* 7, 1 (2022), 13.

[30] Joanna Goode, Kirsten Peterson, Joyce Malyn-Smith, and Gail Chapman. 2020. Online professional development for high school computer science teachers: Features that support an equity-based professional learning community. *Computing in Science & Engineering* 22, 5 (2020), 51–59.

[31] Marc Goutier, Christopher Diebel, Martin Adam, and Alexander Benlian. 2024. Proactive and Reactive Help from Intelligent Agents in Identity-Relevant Tasks. (2024).

[32] Carolin Graßmann and Carsten C Schermuly. 2021. Coaching with artificial intelligence: Concepts and capabilities. *Human Resource Development Review* 20, 1 (2021), 106–126.

[33] Foteini Grivokostopoulou, Konstantinos Kovas, and Isidoros Perikos. 2020. The effectiveness of embodied pedagogical agents and their impact on students learning in virtual worlds. *Applied Sciences* 10, 5 (2020), 1739.

[34] Fernando Gutierrez and John Atkinson. 2011. Adaptive feedback selection for intelligent tutoring systems. *Expert Systems with Applications* 38, 5 (2011), 6146–6152.

[35] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[36] Yann Hicke, Jadon Geathers, Niroop Rajashekar, Colleen Chan, Anyanate Gwendolyne Jack, Justin Sewell, Mackenzi Preston, Susannah Cornes, Dennis Shung, and Rene Kizilcec. 2025. MedSimAI: simulation and formative feedback generation to enhance deliberate practice in medical education. *arXiv preprint arXiv:2503.05793* (2025).

[37] Garron Hillaire, Rick Waldron, Joshua Littenberg-Tobias, Meredith Thompson, Sara O'Brien, Gabrielle R Marvez, and Justin Reich. 2022. Digital clinical simulation suite: Specifications and architecture for simulation-based pedagogy at

scale. In *Proceedings of the Ninth ACM Conference on Learning@ Scale*. 212–221.
[38] Amruta Hingmire, Rushikesh Mashalkar, Shrinand Mathpati, Gajanan Kulkarni, and Girish Nagarkar. 2024. Revolutionary AI-Based Fitness Trainer: A Comprehensive Review. In *International Conference on ICT for Sustainable Development*. Springer, 537–547.
[39] Friederike Holderried, Christian Stegemann-Philipps, Anne Herrmann-Werner, Teresa Festl-Wietek, Martin Holderried, Carsten Eickhoff, Moritz Mahling, et al. 2024. A language model–powered simulated patient with automated feedback for history taking: Prospective study. *JMIR Medical Education* 10, 1 (2024), e59213.
[40] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 159–166.
[41] Qiaolei Jiang, Yadi Zhang, and Wenjing Pian. 2022. Chatbot as an emergency exist: Mediated empathy for resilience via human-AI interaction during the COVID-19 pandemic. *Information processing & management* 59, 6 (2022), 103074.
[42] Brigitte Jordan and Austin Henderson. 1995. Interaction analysis: Foundations and practice. *The journal of the learning sciences* 4, 1 (1995), 39–103.
[43] Kangsoo Kim, Luke Boelling, Steffen Haesler, Jeremy Bailenson, Gerd Bruder, and Greg F Welch. 2018. Does a digital assistant need a body? The influence of visual embodiment and social behavior on the perception of intelligent virtual agents in AR. In *2018 IEEE international symposium on mixed and augmented reality (ISMAR)*. IEEE, 105–114.
[44] Ekaterina Kochmar, Dung Do Vu, Robert Belfer, Varun Gupta, Iulian Vlad Serban, and Joelle Pineau. 2020. Automated personalized feedback improves learning gains in an intelligent tutoring system. In *International conference on artificial intelligence in education*. Springer, 140–146.
[45] Kenneth R Koedinger, Julie L Booth, and David Klahr. 2013. Instructional complexity and the science to constrain it. *Science* 342, 6161 (2013), 935–937.
[46] Ha-Kyung Kong, Rachel Lowy, Youjin Choi, and Jennifer G Kim. 2025. Working Together Toward Interdependence: Chatbot-Based Support for Balanced Social Interactions Between Neurodivergent and Neurotypical Individuals. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.
[47] Konstantinos Koukourikos, Areti Tsaloglidou, Lambrini Kourkouta, Ioanna V Papathanasiou, Christos Iliadis, Aikaterini Fratzana, and Aspasia Panagiotou. 2021. Simulation in clinical nursing education. *Acta Informatica Medica* 29, 1 (2021), 15.
[48] Matthias Kraus, Diana Betancourt, and Wolfgang Minker. 2023. Does it affect you? social and learning implications of using cognitive-affective state recognition for proactive human-robot tutoring. In *2023 32nd IEEE international conference on robot and human interactive communication (RO-MAN)*. IEEE, 928–935.
[49] Matthias Kraus, Nicolas Wagner, Ron Riekenbrauck, and Wolfgang Minker. 2023. Improving proactive dialog agents using socially-aware reinforcement learning. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. 146–155.
[50] Prashant Kumar and Susan Somerville. 2024. Exploring in-person self-led debriefings for groups of learners in simulation-based education: an integrative review. *Advances in Simulation* 9, 1 (2024), 5.
[51] Geonsun Lee, Dae Yeol Lee, Guan-Ming Su, and Dinesh Manocha. 2024. "may i speak?": Multi-modal attention guidance in social vr group conversations. *IEEE Transactions on Visualization and Computer Graphics* 30, 5 (2024), 2287–2297.
[52] Geonsun Lee, Min Xia, Nels Numan, Xun Qian, David Li, Yanhe Chen, Achin Kulshrestha, Ishan Chatterjee, Yinda Zhang, Dinesh Manocha, et al. 2025. Sensible Agent: A Framework for Unobtrusive Interaction with Proactive AR Agents. (2025).
[53] Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S Rodriguez, and Jon E Froehlich. 2024. GazePointAR: A context-aware multimodal voice assistant for pronoun disambiguation in wearable augmented reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–20.
[54] Chenyi Li, Guande Wu, Gromit Yeuk-Yin Chan, Dishita Gdi Turakhia, Sonia Castelo Quispe, Dong Li, Leslie Welch, Claudio Silva, and Jing Qian. 2025. Satori: Towards Proactive AR Assistant with Belief-Desire-Intention User Modeling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–24.
[55] Yahong Lian, Chunyao Song, and Tingjian Ge. 2025. ITMPRec: Intention-based Targeted Multi-round Proactive Recommendation. In *Proceedings of the ACM on Web Conference 2025*. 4171–4182.
[56] Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive conversational agents in the post-chatgpt world. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*. 3452–3455.
[57] Xingyu Bruce Liu, Shitao Fang, Weiyan Shi, Chien-Sheng Wu, Takeo Igarashi, and Xiang'Anthony' Chen. 2025. Proactive conversational agents with inner thoughts. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.
[58] Ziyi Liu, Zhengzhe Zhu, Lijun Zhu, Enze Jiang, Xiyun Hu, Kylie A Peppler, and Karthik Ramani. 2024. Classmeta: Designing interactive virtual classmate to promote vr classroom participation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.
[59] Yaxi Lu, Shenzhi Yang, Cheng Qian, Guirong Chen, Qinyu Luo, Yesai Wu, Huadong Wang, Xin Cong, Zhong Zhang, Yankai Lin, et al. 2024. Proactive agent: Shifting llm agents from reactive responses to active assistance. *arXiv preprint arXiv:2410.12361* (2024).
[60] Xueming Luo, Marco Shaojun Qin, Zheng Fang, and Zhe Qu. 2021. Artificial intelligence coaches for sales agents: Caveats and solutions. *Journal of Marketing* 85, 2 (2021), 14–32.
[61] Subhankar Maity and Aniket Deroy. 2024. Generative ai and its impact on personalized intelligent tutoring systems. *arXiv preprint arXiv:2410.10650* (2024).
[62] Nicole D Martin, Allen Antoine, Andrea Wilson Vazquez, and Cydny Black. 2024. Training effective facilitators to scale equity-focused computer science professional learning. In *Proceedings of the 2024 on RESPECT Annual Conference*. 270–277.
[63] Edward McAuley, Terry Duncan, and Vance V Tammen. 1989. Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research quarterly for exercise and sport* 60, 1 (1989), 48–58.
[64] Michelle McLean. 2003. What can we learn from facilitator and student perceptions of facilitation skills and roles in the first year of a problem-based learning curriculum? *BMC Medical Education* 3, 1 (2003), 9.
[65] Kelly Minotti, Guillaume Loup, Thibault Harquin, and Samir Otmane. 2024. Exploring Immersive Debriefing in Virtual Reality Training: A Comparative Study. In *Proceedings of the 30th ACM Symposium on Virtual Reality Software and Technology*. 1–10.
[66] Ryo Nakahashi and Seiji Yamada. 2021. Balancing performance and human autonomy with implicit guidance agent. *Frontiers in artificial intelligence* 4 (2021), 736321.
[67] Mohit Nayak, Jari Kangas, and Roope Raisamo. 2025. A Study of NLP-Based Speech Interfaces in Medical Virtual Reality. *Multimodal Technologies and Interaction* 9, 6 (2025), 50.
[68] Tatsuya Oyama, Chihiro Sasaki, Chika Oshima, and Koichi Nakayama. 2021. AI Facilitator Allows Participants to Conduct a Friendly Discussion and Contribute to Feasible Proposals. In *International Conference on Human-Computer Interaction*. Springer, 523–530.
[69] Heather L O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (2018), 28–39.
[70] Veljko Pejovic and Mirco Musolesi. 2015. Anticipatory mobile computing: A survey of the state of the art and research challenges. *ACM Computing Surveys (CSUR)* 47, 3 (2015), 1–29.
[71] Thanawit Prasongpongchai, Pat Pataranutaporn, Auttasak Lapapirojn, Chonnipa Kanapornchai, Joanne Leong, Pichayoot Ouppaphan, Kavin Winson, Monchai Lertsutthiwong, and Pattie Maes. 2024. Effects of Proactive Interaction and Instructor Choice in AI-Generated Virtual Instructors for Financial Education. In *2024 IEEE Frontiers in Education Conference (FIE)*. IEEE, 1–9.
[72] Thanawit Prasongpongchai, Pat Pataranutaporn, Monchai Lertsutthiwong, and Pattie Maes. 2025. Talk to the Hand: an LLM-powered Chatbot with Visual Pointer as Proactive Companion for On-Screen Tasks. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–16.
[73] Kevin Pu, Daniel Lazaro, Ian Arawjo, Haijun Xia, Ziang Xiao, Tovi Grossman, and Yan Chen. 2025. Assistance or disruption? exploring and evaluating the design and trade-offs of proactive ai programming support. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–21.
[74] Mashfiqui Rabbi, Min Hane Aung, Mi Zhang, and Tanzeem Choudhury. 2015. MyBehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 707–718.
[75] Harish Chaandar Ravichandar, Avnish Kumar, and Ashwin Dani. 2018. Gaze and motion information fusion for human intention inference. *International Journal of Intelligent Robotics and Applications* 2, 2 (2018), 136–148.
[76] Florian Recker, Ricarda Neubauer, Yi Dong, Anna Maria Gschmack, Christian Jenssen, Kathleen Möller, Michael Blaivas, Prats Michael Ignacio, Claudia Lucius, Johannes Ruppert, et al. 2024. Exploring the dynamics of ultrasound training in medical education: current trends, debates, and approaches to didactics and hands-on learning. *BMC Medical Education* 24, 1 (2024), 1311.
[77] Etienne Rivière, Danielle Saucier, Alexandre Lafleur, Miriam Lacasse, and Gilles Chiniara. 2018. Twelve tips for efficient procedural simulation. *Medical Teacher* 40, 7 (2018), 743–751.
[78] Andrei Romancenco, Iurie Săratilă, Ivan Ababiy, Gheorghe Rojnoveanu, Otilia Dandara, and Larisa Spinei. 2024. Bridging theory and practice: enhancing medical education through simulation-based training methods. *Revista de Ştiinţe ale Sănătăţii din Moldova* 2 (2024), 68–73.
[79] Lauren A Sattler, Chad Schuety, Mark Nau, Daniel V Foster, John Hunninghake, Tyson Sjulin, and Joshua Boster. 2020. Simulation-based medical education improves procedural confidence in core invasive procedures for military internal medicine residents. *Cureus* 12, 12 (2020).

[80] Taylor Sawyer, Marjorie White, Pavan Zaveri, Todd Chang, Anne Ades, Heather French, JoDee Anderson, Marc Auerbach, Lindsay Johnston, and David Kessler. 2015. Learn, see, practice, prove, do, maintain: an evidence-based pedagogical framework for procedural skill training in medicine. *Academic Medicine* 90, 8 (2015), 1025–1033.

[81] Benedikt Schmidt, Sebastian Benchea, Rüdiger Eichin, and Christian Meurisch. 2015. Fitness tracker or digital personal coach: how to personalize training. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. 1063–1067.

[82] Philipp M Scholl, Matthias Wille, and Kristof Van Laerhoven. 2015. Wearables in the wet lab: a laboratory system for capturing and guiding experiments. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 589–599.

[83] Sangwon Seo, Lauren R Kennedy-Metz, Marco A Zenati, Julie A Shah, Roger D Dias, and Vaibhav V Unhelkar. 2021. Towards an AI coach to infer team mental model alignment in healthcare. In *2021 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. IEEE, 39–44.

[84] Aidin Shahrezaei, Maryam Sohani, Soroush Taherkhani, and Seyed Yahya Zarghami. 2024. The impact of surgical simulation and training technologies on general surgery education. *BMC Medical Education* 24, 1 (2024), 1297.

[85] Mark Sherer, James E Maddux, Blaise Mercandante, Steven Prentice-Dunn, Beth Jacobs, and Ronald W Rogers. 1982. The self-efficacy scale: Construction and validation. *Psychological reports* 51, 2 (1982), 663–671.

[86] Matt Sibbald, Kyla Caners, and Bingxian Wang. 2019. Managing cognitive load in simulations: exploring the role of simulation technologists. *Canadian Medical Education Journal* 10, 4 (2019), e48.

[87] Maninder Singh and Andrew Restivo. 2023. Task trainers in procedural skills acquisition in medical simulation. In *StatPearls [Internet]*. StatPearls Publishing.

[88] Olle Ten Cate, Natasha Khursigara-Slattery, Richard L Cruess, Stanley J Hamstra, Yvonne Steinert, and Robert Sternszus. 2024. Medical competence as a multilayered construct. *Medical Education* 58, 1 (2024), 93–104.

[89] Danielle R Thomas, Jionghao Lin, Erin Gatz, Ashish Gurung, Shivang Gupta, Kole Norberg, Stephen E Fancsali, Vincent Aleven, Lee Branstetter, Emma Brunskill, et al. 2024. Improving student learning with hybrid human-AI tutoring: A three-study quasi-experimental investigation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. 404–415.

[90] Maria Suong Tjønnås, Anita Das, Cecilie Våpenstad, and Solveig Osborg Ose. 2022. Simulation-based skills training: a qualitative interview study exploring surgical trainees' experience of stress. *Advances in Simulation* 7, 1 (2022), 33.

[91] Jakob Carl Uhl, Rodrigo Gutierrez, Georg Regal, Helmut Schrom-Feiertag, Benjamin Schuster, and Manfred Tscheligi. 2024. Choosing the right reality: A comparative analysis of tangibility in immersive trauma simulations. In *Proceedings of the 2024 chi conference on human factors in computing systems*. 1–17.

[92] Jakob Carl Uhl, Helmut Schrom-Feiertag, Georg Regal, Katja Gallhuber, and Manfred Tscheligi. 2023. Tangible immersive trauma simulation: is mixed reality the next level of medical skills training?. In *Proceedings of the 2023 chi conference on human factors in computing systems*. 1–17.

[93] Aldo Uriarte-Portillo, Ramón Zatarain-Cabada, María Lucía Barrón-Estrada, María Blanca Ibáñez, and Lucía-Margarita González-Barrón. 2023. Intelligent augmented reality for learning geometry. *Information* 14, 4 (2023), 245.

[94] Pedro Valente. 2025. Integrating Proactive Conversational Agents in Hybrid User Interfaces: A Task-Based Analysis. In *Companion Proceedings of the 30th International Conference on Intelligent User Interfaces*. 222–224.

[95] Joost van Oijen. 2024. Augmented Intelligence for Instructional Systems in Simulation-Based Training. In *International Conference on Human-Computer Interaction*. Springer, 89–101.

[96] Juan C Vassallo, Bárbara Gouguenheim, Analía Ghiglione, Nélida Bravo, Carla I Prudencio, Florencia Villois, Yamila Abadie, Ana Zubieta, Carol Golini, Victoria Villar, et al. 2015. Lumbar puncture training using simulation-based educational strategies. Experience in a clinical pediatric residency. *Arch Argent Pediatr* 113, 6 (2015), 544–549.

[97] Margaret Verkuyl, Efrem Violato, Theresa Southam, Mélanie Lavoie-Tremblay, Sandra Goldsworthy, Diane MacEachern, and Lynda Atack. 2024. Facilitators' experiences with virtual simulation and their impact on learning. *Advances in Simulation* 9, 1 (2024), 54.

[98] Amy Wang, Roma Ruparel, Anna Iurchenko, Paul Jhun, Julie Anne Séguin, Patricia Strachan, Renee Wong, Alan Karthikesalingam, Yossi Matias, Avinatan Hassidim, et al. 2025. Generative AI for medical education: Insights from a case study with medical students and an AI tutor for clinical reasoning. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[99] Duo Wang, Kyrian Liang, Wenxuan Song, Qingxiao Zheng, Zyra Sheikh, Jen Whiting, Victor Lu, Mike Yao, and Caroline Cao. 2026. EmSim: AI-Driven VR Simulation for De-Escalation Training in Law Enforcement. In *2026 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*. IEEE.

[100] Duo Wang, Jianwei Ni, Yuhan Zhou, Weiyu Ding, Wenxuan Song, Shandra Jamison, Mike Yao, and Caroline Cao. 2026. SPATIALTUTOR: Object-Aware Mixed Reality Training for Procedural Medical Skills Training with AI-Driven Support. In *2026 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*. IEEE.

[101] Duo Wang, Xiwei Wang, Qingxiao Zheng, Bingxin Tao, and Guomeng Zheng. 2020. How interaction paradigms affect user experience and perceived interactivity in virtual reality environment. In *International Conference on Human-Computer Interaction*. Springer, 223–234.

[102] Isaac Wang, Jesse Smith, and Jaime Ruiz. 2019. Exploring virtual agents for augmented reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[103] Giles Westerfield, Antonija Mitrovic, and Mark Billinghurst. 2015. Intelligent augmented reality training for motherboard assembly. *International Journal of Artificial Intelligence in Education* 25, 1 (2015), 157–172.

[104] Qingming Wu, Yubin Wang, Lili Lu, Yong Chen, Hui Long, and Jun Wang. 2022. Virtual simulation in undergraduate medical education: a scoping review of recent practice. *Frontiers in medicine* 9 (2022), 855403.

[105] Shuchang Xu, Chang Chen, Zichen Liu, Xiaofu Jin, Lin-Ping Yuan, Yukang Yan, and Huamin Qu. 2024. Memory Reviver: Supporting Photo-Collection Reminiscence for People with Visual Impairment via a Proactive Chatbot. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–17.

[106] Recai Yilmaz, Mohamad Bakhaidar, Ahmad Alsayegh, Nour Abou Hamdan, Ali M Fazlollahi, Trisha Tee, Ian Langleben, Alexander Winkler-Schwartz, Denis Laroche, Carlo Santaguida, et al. 2024. Real-Time multifaceted artificial intelligence vs In-Person instruction in teaching surgical technical skills: a randomized controlled trial. *Scientific Reports* 14, 1 (2024), 15130.

[107] Meriem Zerkouk, Miloud Mihoubi, and Belkacem Chikhaoui. 2025. A Comprehensive Review of AI-based Intelligent Tutoring Systems: Applications and Challenges. *arXiv preprint arXiv:2507.18882* (2025).

[108] Tianyi Zhang, Colin Au Yeung, Emily Aurelia, Yuki Onishi, Neil Chulpongsatorn, Jiannan Li, and Anthony Tang. 2025. Prompting an Embodied AI Agent: How Embodiment and Multimodal Signaling Affects Prompting Behaviour. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–25.

[109] Sebastian Zhao, Alan Zhu, Hussein Mozannar, David Sontag, Ameet Talwalkar, and Valerie Chen. 2025. Codinggenie: A proactive llm-powered programming assistant. In *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering*. 1168–1172.

[110] Qingxiao Zheng, Yiliu Tang, Yiren Liu, Weizi Liu, and Yun Huang. 2022. UX research on conversational human-AI interaction: A literature review of the ACM digital library. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–24.

[111] Youness Zidoun and Abdelmoniem EL Mardi. 2024. Artificial Intelligence (AI)-Based simulators versus simulated patients in undergraduate programs: A protocol for a randomized controlled trial. *BMC Medical Education* 24, 1 (2024), 1260.