

CAESAR: Mixed Reality Search-and-Rescue Platform for Cognitive Attack Evaluation

Mingyu Zhu , Jiangong Chen , Yu Liu , Hillmer Chona , Zifan Zhou , Jianan Jiang , Jonathan Dodge, Feng Qian , Mahdi Imani , Tian Lan , Gang Tan , and Bin Li 

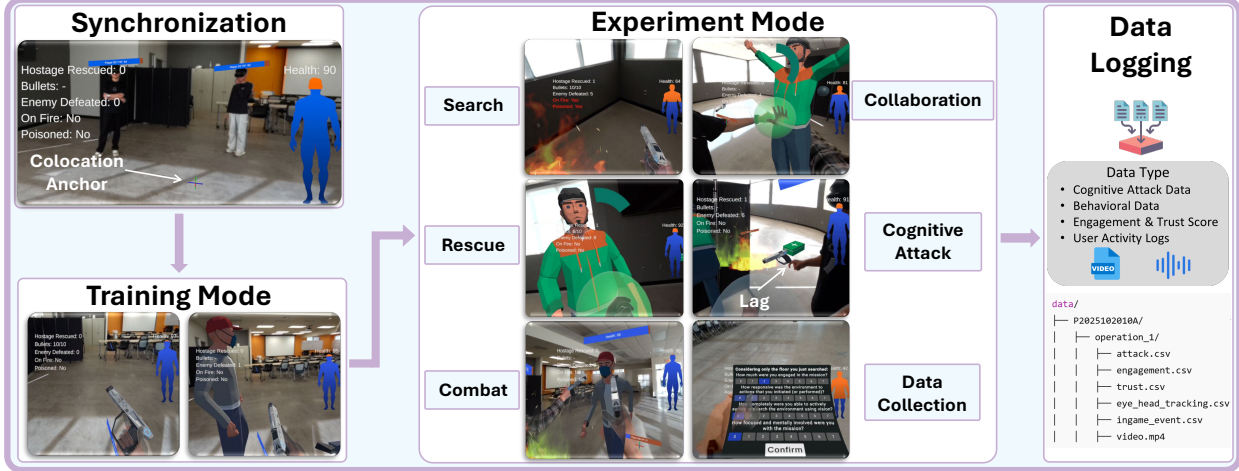


Figure 1: Overview of the mixed reality search-and-rescue testbed. The platform begins with synchronization using a colocation system, followed by a training phase, and then transitions into experimental tasks including search, rescue, and combat. Collaborative interactions, cognitive attacks, and data collection are seamlessly integrated into the testbed, supporting synchronized multi-user scenarios, structured training, and systematic evaluation under cognitive attacks.

Abstract— Mixed Reality (MR) enables immersive training by tightly coupling virtual content with the physical environment and supporting embodied interaction. In safety-critical collaborative tasks such as search-and-rescue, however, users rely heavily on virtual elements for perception, coordination, and decision-making, making them vulnerable to cognitive attacks that subtly distort visual feedback, virtual content, or system synchronization. While prior MR security research has examined individual attack mechanisms, evaluations are typically conducted in isolated or single-user settings, and there is a lack of standardized collaborative testbeds for systematic adversarial experimentation. In this paper, we develop CAESAR, an end-to-end multi-user MR framework for systematic evaluation of cognitive attacks in collaborative search-and-rescue training. CAESAR features a modular architecture that integrates four classes of cognitive attacks within a synchronized collaborative scenario and incorporates a multimodal data collection pipeline that captures behavioral, performance, and subjective measures under adversarial conditions. We conduct a controlled user study with an in-depth analysis of the latency-based attack, demonstrating measurable effects on task performance and user behavior while characterizing users’ subjective experience under adversarial conditions.

Index Terms—Mixed reality, cognitive attack, search and rescue, evaluation testbed

1 INTRODUCTION

Mixed Reality (MR) enables immersive training experiences that tightly couple virtual content with the physical environment, allowing users to engage in embodied interactions with both digital and real-world elements [8, 11, 12, 30] while gaining hands-on experience that is critical

for effective skill acquisition [2, 7, 14]. Unlike Virtual Reality (VR), which fully replaces the physical world with synthetic simulations [24, 45], MR maintains physical grounding while augmenting perception and interaction, enabling training that more closely reflects real-world operations. These capabilities are particularly valuable in safety-critical domains, where real-world training may be dangerous or infeasible, such as firefighter training [7, 32], emergency medical procedures [2], and collaborative search-and-rescue operations [23, 36, 39].

In collaborative search-and-rescue MR training, virtual elements directly shape users’ attention, perception, and coordinated action across shared tasks. Because decision-making depends on accurate spatial alignment, synchronized state representation, and interpretation of virtual content, even subtle perturbations can influence task execution and collective situational awareness. This tight coupling between perception, interaction, and system infrastructure introduces a vulnerability: perturbations to visual feedback, virtual content, user-state information, or networked synchronization can distort cognition without immediately alerting users. Despite this risk, existing search-and-rescue MR systems primarily aim to enhance operational effectiveness, realism,

- Mingyu Zhu, Jiangong Chen, Hillmer Chona, Zifan Zhou, Jianan Jiang, Jonathan Dodge, Gang Tan, and Bin Li are with The Pennsylvania State University. E-mail: {mintrey, jiangong, hac5324, zqz5454, jnjiang, jxd6067, gtan, binli}@psu.edu.
- Yu Liu and Feng Qian are with the University of Southern California. E-mail: {yliu2170, fengqian}@usc.edu.
- Mahdi Imani is with Northeastern University. E-mail: m.imani@northeastern.edu.
- Tian Lan is with The George Washington University. E-mail: tlan@email.gwu.edu.

Manuscript received xx xxx. 202x; accepted xx xxx. 202x. Date of Publication xx xxx. 202x; date of current version xx xxx. 202x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.202x.xxxxxxx

and situational awareness. For example, prior work has explored agent-based MR simulation for combat search-and-rescue command and control [21], collaborative perception systems [37], and Augmented Reality (AR)-based urban search-and-rescue support integrating AI and localization [38]. However, these systems are not designed to support controlled injection of adversarial perturbations or systematic evaluation of user responses under degraded or manipulated conditions.

Prior MR security research [6, 33, 41] has examined individual attack mechanisms, but typically in isolated or single-user contexts. As a result, there remains a lack of standardized collaborative, end-to-end MR testbeds that enable systematic evaluation of cognitive attacks within realistic multi-user training scenarios. In particular, few platforms capture synchronized behavioral, decision-making, and cognitive data under adversarial conditions. Such data are essential for understanding user behavior, performance, and cognitive responses and for advancing cognitive security research [29].

To address this gap, we develop CAESAR (Cognitive Attack Evaluation in Search-And-Rescue), an end-to-end collaborative MR experimental framework and testbed for systematic evaluation of cognitive attacks in multi-user search-and-rescue scenarios. CAESAR integrates four classes of cognitive attacks, including MR display perturbation, MR content manipulation, MR user status breach, and network and compute perturbation, within a synchronized end-to-end scenario. Integrated with a multimodal data collection pipeline that records attack events, behavioral signals, task performance, and subjective measures, the testbed enables comprehensive analysis of user behavior and collaborative dynamics under adversarial conditions.

The CAESAR testbed models a realistic multi-user search-and-rescue mission in which participants locate and rescue hostages within a hazardous and adversarial MR environment. The scenario incorporates a unified health system that introduces time pressure and risk-reward tradeoffs, dynamic environmental hazards that degrade perception and safety, and AI-driven adversaries that impose continuous external threats. Hostage rescue requires sustained interaction and coordination, coupling combat and rescue objectives through interdependent task mechanics. The platform maintains fully synchronized multi-user collaboration to preserve shared situational awareness and coordinated action. To scale mission complexity within constrained physical spaces, CAESAR employs a virtual elevator mechanism that structures the experience into multi-stage missions across distinct virtual environments.

These design elements establish CAESAR as an end-to-end collaborative MR testbed for systematic evaluation of cognitive attacks in search-and-rescue training. Our contributions are as follows:

- We develop and open-source¹, to our knowledge, the first end-to-end multi-user MR testbed designed for systematic evaluation of cognitive attacks in collaborative search-and-rescue scenarios.
- We implement four classes of cognitive attacks within the testbed and conduct an in-depth user study that uses latency-based display perturbation to validate the framework, demonstrating its impact on task performance, user behavior, and subjective experience.
- We provide an integrated experimental framework that includes configurable gameplay modes, controlled attack injection, and a real-time multimodal data collection pipeline, enabling systematic user studies and generating rich datasets for post-hoc analysis, and supporting future collaborative MR cognitive attack research across diverse application scenarios.

2 RELATED WORK

2.1 Security Attacks and Recognizability in MR

The integration of real-world environments with virtual content in MR systems introduces novel security risks and attack surfaces. Attackers can manipulate visual information in MR to present misleading content, potentially causing users to make unsafe or erroneous decisions. Virtual objects may also occlude real-world elements or distort spatial relationships, further increasing the risk of user error. Recent work has begun to formalize the broader landscape of MR cognitive

and deception attacks. For example, Teymourian et al. [35] proposed a comprehensive ontology of MR deception attacks spanning security, cognition, and information-theoretic perspectives. Alongside this emerging understanding, prior research has investigated different types of attacks in MR and AR systems, with a focus on their recognizability by users. Cheng et al. [6] conducted an early empirical study of perceptual manipulation attacks in MR, showing that users often struggle to attribute anomalous perceptions to malicious interference. Subsequent work by Cheng et al. [5] demonstrated that embedding users within AR interfaces undermines traditional UI security assumptions, making attacks harder to detect and reason about. Other studies [41–43] examined visual information attacks such as misleading overlays, occlusion of critical elements, and misaligned virtual content. Sajid et al. [33] explored interaction attacks in MR and found that manipulations of spatial cues are particularly difficult for users to recognize.

While these studies provide valuable insights into individual attack mechanisms, they are typically evaluated in isolation and task-specific or single-user settings. As a result, they offer limited insight into how MR security threats manifest in realistic, collaborative environments. As multi-user MR applications become increasingly common, how attacks affect shared spatial understanding, coordination, and collective decision-making remains underexplored.

2.2 Co-located Multi-user Collaboration in MR

Many MR applications, especially those for training, planning, and decision support, are inherently collaborative, which fundamentally alters how security threats manifest and propagate across users. MR systems support both co-located [4, 18, 20] and remote [16, 34, 44] collaboration, enabling users to share virtual content while maintaining awareness of the physical environment and others' actions. For co-located users, MR facilitates situated and embodied collaboration by anchoring virtual objects to a shared physical space. Prior work has shown that effective multi-user MR collaboration depends on maintaining a common spatial reference frame and supporting awareness of collaborators' activities [1, 10]. Several systems have been proposed to enhance shared understanding and interaction consistency. For example, Gottsacker et al. [15] presented a decoupled rendering approach that improves communication by visualizing collaborators' actions from their respective viewpoints, while Miria [3] provides a co-located MR data visualization toolkit based on manual coordinate synchronization.

Although these systems improve usability and scalability, they are primarily designed to enhance interaction quality rather than assess system robustness or security. From a security perspective, inconsistencies in synchronization, misaligned spatial representations, or asymmetric visual states across users can enable attacks that are difficult to detect individually. However, few existing collaborative MR frameworks support systematic evaluation of such vulnerabilities under adversarial or degraded conditions.

2.3 Simulation and Training in MR

MR and VR technologies have been widely adopted for simulation and training because they provide immersive, interactive experiences while reducing the cost and risk of real-world training. They have been applied across domains such as education, medical training, firefighting, and search and rescue. For example, Zhao et al. [45] demonstrated that VR video and pre-captured scenes can substitute for physical field trips in geometry education, while Lorenz et al. [24] incorporated haptic feedback into VR-based hip surgery training to improve procedural realism. Compared with VR, MR enables training directly within real-world environments while augmenting perception and interaction, making it particularly suitable for safety-critical, team-based tasks. For example, Dong et al. [7] introduced a collaborative MR system for firefighter training that emphasizes synchronized multi-user experiences. Search and rescue operations are similarly safety-critical and inherently collaborative, involving complex social and cognitive dynamics [23, 36, 39]. Ji et al. [21] presented an early agent-based MR simulation framework for combat search and rescue command and control, while Ajina [37] improves situational awareness in search and rescue through shared perception, focusing on operational effectiveness

¹<https://github.com/SNEC-Lab-PSU/CAESAR>

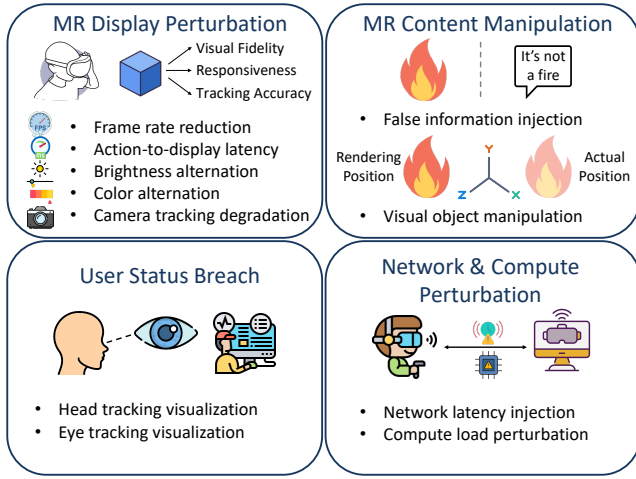


Figure 2: Four classes of cognitive attacks in MR environments.

rather than security or adversarial robustness. Related work has also explored AR-based approaches for urban search and rescue, such as integrating AI and localization to support human-canine teams [38]. Beyond application-specific training systems, research platforms such as ILLIXR [19] provide modular XR infrastructure for end-to-end systems experimentation. Rather than targeting a specific application domain, these platforms support XR systems research and development.

Despite these advances, prior MR and VR training systems primarily aim to improve realism, usability, and task performance. Few are designed as experimental testbeds for evaluating security threats, and they do not support controlled injection of cognitive attacks, systematic variation of attack parameters, or unified measurement of system- and user-level responses in multi-user settings. As a result, how collaborative MR training systems and their users behave under security threats remains poorly understood. In contrast, we present an end-to-end collaborative MR experimental framework designed as a testbed for systematic evaluation of cognitive attacks, integrating configurable attack injection, synchronized multi-user experimentation, and multi-modal behavioral measurement within collaborative search-and-rescue scenarios. By doing so, we bridge the gap between MR security research and MR-based training systems, contributing toward safer and more resilient multi-user MR platforms.

3 COGNITIVE ATTACK

Modern MR systems, while enabling immersive collaboration, also introduce opportunities for adversarial manipulation of user cognition. We define cognitive attacks as intentional disruptions that exploit MR’s access to sensory, perceptual, and behavioral modalities to overload, mislead, or destabilize users during task execution. In this section, we present the cognitive attack classes implemented within our framework for systematic evaluation in collaborative MR environments. These attacks leverage critical MR system dependencies, such as visual feedback, sensor data, or network and compute resources, to expose potential vulnerabilities and their impact on user cognition. Figure 2 shows an overview of the attack classes considered in this work.

3.1 Threat Model and Design Rationale

Threat Model. Our objective is to study the effects of cognitive attacks on user behavior and task performance, rather than the feasibility of compromising MR platforms or bypassing platform security mechanisms. Accordingly, we consider an adversary whose goal is to manipulate user cognition in MR environments by degrading perception, misleading interpretation, or increasing cognitive load during task execution, rather than causing system failure. The attacker does not require physical access to the victim but can influence the MR system through software, networking, or content-level channels. Such an attacker may take the form of a compromised MR application or plugin [5], a malicious or misconfigured server in a client-server architecture [33], or

a third-party service with access to MR system APIs [6]. We assume the attacker cannot bypass hardware-level isolation or the underlying operating system but operates at the application or middleware level by exploiting exposed interfaces and trusted execution paths commonly used in collaborative MR systems. Under this model, the attacker can modify or delay application-level events, alter virtual content or feedback logic, access user sensing streams that are legitimately exposed through platform APIs or granted permissions, and interfere with network synchronization or compute resource allocation.

Design Rationale. Modern MR experiences rely on a small set of critical system dependencies that directly shape user cognition, including (1) continuous visual feedback and action–feedback coupling, (2) semantically meaningful virtual content and interaction logic, (3) rich user sensing pipelines that capture behavioral signals, and (4) networked and compute-intensive execution for multi-user synchronization and real-time processing. Attacks that target any of these dependencies can systematically disrupt perception, attention, trust, or decision-making without immediately alerting the user.

Based on this observation, we organize the cognitive attacks into four categories corresponding to these core MR system dependencies: MR display perturbation, MR content manipulation, MR user status breach, and network and compute perturbation. Unlike ontology-oriented frameworks that seek to comprehensively characterize MR deception attacks [35], our implementation-oriented categorization is designed to support systematic deployment and evaluation within collaborative MR systems instead of proposing a comprehensive attack taxonomy. Rather than enumerating all exploits, these categories reflect distinct security-relevant MR system layers identified in prior work [9, 13, 17], capturing common and high-impact cognitive attack surfaces. They provide structured coverage of adversarial manipulation across perception, interaction logic, sensing pipelines, and infrastructure, motivating the implemented attack classes below.

3.2 Implemented Cognitive Attack Classes

MR Display Perturbation. MR display perturbation represents a class of attacks that target continuous visual feedback and action–feedback coupling in MR environments. An attacker operating at the application or rendering layer can intentionally degrade visual fidelity, responsiveness, or tracking stability without causing overt system failure. Such manipulations disrupt the tight perception–action loop, forcing users to adapt under degraded conditions and increasing cognitive workload. The following attack instances demonstrate practical realizations of display perturbation in MR systems and are implemented in our testbed.

Frame Rate Reduction. Artificial delays can be introduced into the update or rendering loop, reducing the effective frame rate perceived by the user. Such degradation can be achieved through inserted wait states, blocking operations, or compute-intensive callbacks. In our testbed, we implement this attack by injecting controlled delays into the main update loop. Reduced frame rates disrupt visual smoothness and weaken action–feedback coupling, leading to delayed interaction.

Action-to-Display Latency. Delays can be introduced between user input events and corresponding visual updates, disrupting the temporal consistency of the action–feedback loop. This can be realized by intercepting or delaying event handling callbacks at the application layer. In our testbed, we implement this attack by inserting controlled delays and random jitter into input processing before corresponding rendering updates are triggered. The resulting action–response mismatch degrades responsiveness and destabilizes perceived control.

Brightness and Color Alteration. Display-level parameters exposed through platform APIs can be manipulated to alter the visual appearance of the physical environment in MR. Using platform-provided passthrough control interfaces (e.g., Meta’s Passthrough API [26]), brightness and edge color rendering can be adjusted without privileged system access. In our testbed, we implement this attack by dynamically adjusting passthrough brightness and edge color values during task execution. Such alterations distort scene visibility and saliency, potentially misleading perception and increasing cognitive workload.

Camera Tracking Degradation. Tracking stability can be degraded by introducing controlled noise or offsets into pose data consumed

at the application layer. Such perturbations can be applied to camera position and orientation values obtained through platform APIs, without modifying hardware or underlying system components. In our testbed, we implement this attack by injecting stochastic perturbations into the tracked pose data before scene rendering. These disturbances cause scene jitter or drift, weakening spatial alignment and destabilizing perception.

MR Content Manipulation. MR content manipulation represents a class of attacks that target the semantic and interaction logic of virtual objects within the MR environment. An attacker operating at the application layer can modify object properties, interaction handlers, or feedback messages, thereby altering how virtual content is presented and interpreted without causing system failure. Inconsistencies between appearance, behavior, and task semantics can mislead users and degrade decision-making. We implement two content-level attack instances in our testbed, as described below.

False Information Injection. Interaction feedback logic can be altered to provide incorrect or misleading semantic information by modifying system messages or status indicators, without changing the underlying interaction state. In our testbed, we modify the feedback handler associated with a hazardous virtual object to display incorrect information despite its actual behavior. This mismatch between system feedback and environmental state can mislead users and undermine situational awareness.

Visual Object Manipulation. The spatial alignment between a virtual object’s visual representation and its interaction boundary can be deliberately altered. By modifying object transforms or interaction colliders, the rendered object can be displaced from its true interaction location. In our testbed, we shift the visual representation of a hazardous object while keeping its original interaction boundary unchanged. This visual–interaction mismatch disrupts perception–action mapping and may confuse users.

MR User Status Breach. MR user status breach represents a class of attacks that exploit sensor data streams exposed to applications in MR systems. An attacker operating at the application layer can access and log behavioral signals such as head motion and eye gaze without modifying hardware or system components. Such access enables the extraction of sensitive cognitive and behavioral information. Moreover, eye gaze and head motion patterns may allow user identification based on behavioral biometrics [28, 40]. In our testbed, we capture real-time gaze and head tracking data, demonstrating that these sensor streams are directly accessible and recordable at the application layer.

Head and Eye Tracking Visualization. The headset continuously exposes head pose and eye tracking data to applications through platform-provided APIs and granted permissions, including 3D position, orientation, and gaze vectors. In our testbed, we access these streams in real time, compute the gaze focal point, and transmit the collected data to a server for subsequent visualization and analysis. This demonstrates that detailed behavioral signals can be extracted and externally processed, potentially enabling the types of inference and identification risks.

Network and Compute Perturbation. Network and compute perturbation represents a class of attacks that target MR systems’ dependence on communication latency and computational resources. An attacker operating at the application or server layer can intentionally introduce delays, inject artificial workload, or manipulate synchronization logic. Such perturbations degrade responsiveness, disrupt multi-user consistency, and weaken temporal coordination in collaborative MR environments. We implement two network- and compute-level attack instances in our testbed.

Network Latency Injection. Collaborative MR systems rely on timely synchronization of shared virtual state across devices. Synchronization messages can be intentionally delayed to introduce communication latency. In our testbed, we modify the server-side synchronization process of shared objects (e.g., via Meta Shared Spatial Anchors API [27]) to introduce controlled delays between object creation and remote rendering. These induced delays disrupt temporal consistency and degrade the responsiveness and coordination of multi-user MR interactions.

Compute Load Perturbation. Compute load perturbation exploits the limited processing resources of MR applications. Adversarial workload

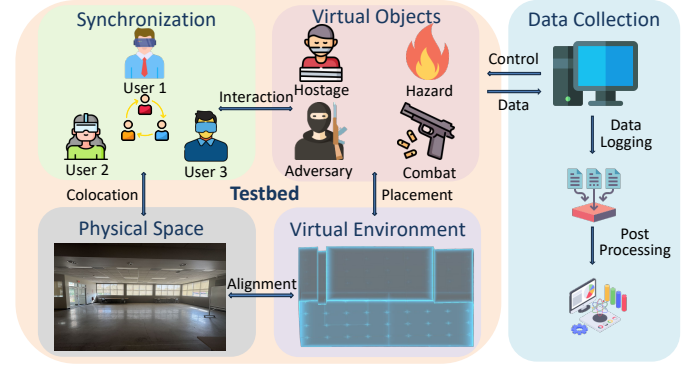


Figure 3: System architecture of the CAESAR testbed. The framework integrates the physical space with the virtual environment through spatial alignment, and synchronizes multiple users via a colocation system. Within the testbed, users interact with virtual objects such as hostages, adversaries, hazards, and weapons to simulate search-and-rescue tasks. A data collection pipeline captures multimodal data for post-hoc analysis.

can be introduced at the application or middleware layer by triggering computation-intensive routines in response to specific events. In our testbed, we inject workload into the execution pipeline to increase CPU usage and reduce rendering responsiveness. The resulting resource contention lowers frame stability and degrades interaction smoothness.

4 SYSTEM OVERVIEW

We designed the CAESAR framework as a multi-user MR testbed for controlled tactical search-and-rescue simulations with integrated support for cognitive attack deployment. As shown in Figure 3, it combines spatially mapped physical environments with persistent virtual objects, synchronized interactions and animations, real-time player-state tracking, and comprehensive data logging.

To bridge physical space and the virtual environment, CAESAR aligns virtual content to the mapped real world using a standard MR spatial mapping pipeline [25]. This alignment enables users to interact naturally with both virtual and physical elements, providing a coherent foundation for immersive simulation. The testbed instantiates mission-critical virtual entities and environmental hazards, which are spatially distributed to form structured search-and-rescue tasks. These configurations integrate operational objectives with environmental constraints to create a realistic collaborative setting. Importantly, the system provides runtime control mechanisms that allow cognitive attacks to be injected during task execution, enabling systematic evaluation of user behavior and performance under adversarial MR conditions.

Within the MR environment, CAESAR maintains consistent state synchronization across all participants to support seamless multi-user collaboration. A colocation system based on a shared spatial reference frame [27] ensures geometric alignment of virtual content across devices, while real-time network synchronization [31] propagates object states, user interactions, animations, and player status information. These mechanisms preserve shared situational awareness and ensure that mission-critical events remain consistent across perspectives.

To support controlled experimental studies, the CAESAR testbed provides a dedicated training mode and configurable experiment modes. The training mode introduces core tasks and interaction mechanisms to familiarize participants with mission procedures. Experiment modes vary task parameters, object distributions, and attack configurations, and allow dynamic runtime switching and cognitive attack injection. The platform also integrates a real-time data collection pipeline that records both system-level and user-level metrics. Captured data includes cognitive attack events, behavioral signals (e.g., head and eye tracking), user activity logs, and subjective measures such as engagement and trust. All data are synchronized and stored on the server for subsequent analysis, enabling fine-grained post-hoc evaluation of user resilience, decision-making, and collaboration under adversarial MR conditions. These components establish CAESAR as a flexible and extensible research framework for systematic evaluation of cognitive

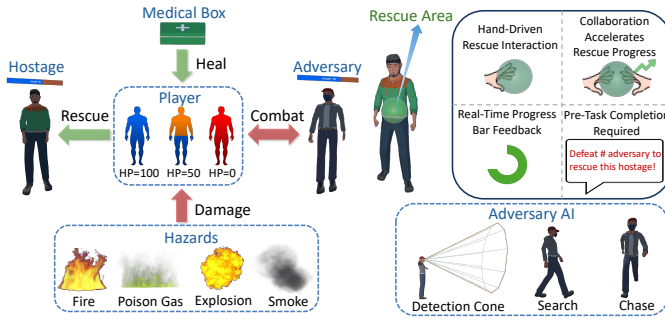


Figure 4: Search-and-rescue features. **(Left):** Unified health system linking hazards, combat, rescue, and recovery. **(Top Right):** Rescue mechanism involving hand-driven interaction, progress feedback, collaboration, and pre-task completion check. **(Bottom Right):** Adversary AI's detect, search, and chase behaviors.

attacks in collaborative MR search-and-rescue environments.

5 SEARCH-AND-RESCUE FEATURES

This section describes the core search-and-rescue features supported by the CAESAR testbed and explains how they shape mission dynamics, user behavior, and experimental control. These components capture key elements of real-world search-and-rescue operations, including time pressure, environmental uncertainty, adversarial interaction, and team coordination, enabling systematic study of collaborative performance under stress.

5.1 Health System

We introduce a unified health system to model survival constraints, time pressure, and resource dependency within the search-and-rescue scenario. Health serves as a shared quantitative state for players, hostages, and adversaries, linking environmental hazards, combat interactions, and rescue urgency within a single framework (see Figure 4 (Left)). Players experience gradual health decay over time and additional damage from environmental hazards and adversarial encounters. Health can only be restored through limited recovery resources, creating strategic trade-offs between exploration, combat, and self-preservation. Hostages also experience health degradation, with varying survival durations that require players to prioritize rescue decisions under uncertainty. Adversaries possess health states that decrease through combat, enabling bidirectional risk during engagement. Importantly, health directly constrains interaction capabilities. When a player's health is depleted, their ability to perform mission actions is restricted until recovery is achieved. This coupling between physiological state and action capability introduces sustained cognitive and behavioral pressure, encouraging risk assessment, coordination, and adaptive decision-making throughout the mission.

5.2 Hostile Environment

We model a hostile environment through four hazard categories: fire, poison gas, explosion, and smoke (see Figure 4 (Left)). These hazards introduce spatial risk, stochastic damage, and perceptual degradation, collectively simulating the uncertainty and dynamics of real-world search-and-rescue operations. Fire and gas impose continuous or residual health loss, while explosions create sudden high-impact events that disrupt navigation and decision-making. Smoke reduces visibility and object detectability, degrading situational awareness. These hazards generate dynamic environmental pressure, forcing players to balance exploration, rescue urgency, and personal safety under uncertain and evolving conditions.

5.3 Rescue Mechanism

The rescue mechanism implements time-sensitive, collaborative interaction under environmental and adversarial pressure (see Figure 4 (Top Right)). Hostages experience gradual health degradation, creating urgency and requiring timely intervention. Successful rescue requires sustained interaction at close range, exposing players to surrounding

risks. To promote teamwork, rescue efficiency increases when multiple players coordinate their interaction, encouraging cooperative behavior rather than isolated action. Rescue attempts are further constrained by mission requirements: players must eliminate a specified number of adversaries before a hostage can be saved. The required number increases progressively across rescues, linking combat performance with rescue progress and introducing increasing coordination demands. These mechanisms transform rescue from a simple interaction into a structured task involving prioritization, synchronization, and risk management under evolving mission constraints.

5.4 Adversary AI and Combat System

Adversaries introduce adaptive threat pressure through autonomous search, detection, pursuit, and attack behaviors (see Figure 4 (Bottom Right)). In the absence of nearby players, adversaries patrol the environment; once a player is detected, they transition to pursuit and cause damage during close-range engagement. This dynamic response creates continuous situational awareness demands and spatial risk for players. Combat interactions are bidirectional: players can defeat adversaries through ranged or close-range engagement, while adversaries can also inflict damage in return. Ranged combat requires ammunition management and timely reloading, introducing resource constraints that influence engagement decisions. Damage exchange is configurable, allowing controlled adjustment of combat intensity for different experimental conditions. Adaptive adversary behavior and resource-constrained combat establish sustained threat exposure, requiring players to balance combat decisions, positioning, and coordination under ongoing threat.

5.5 Elevator System and Visibility Mechanics

To structure mission progression and extend the effective study space, we introduce a virtual elevator system that segments the environment into multiple virtual floors. Each floor contains distinct distributions of hostages, adversaries, and hazards, enabling controlled variation in task difficulty and duration. The distributions are predefined and configurable, enabling reproducible experimental conditions across participants. Transition between floors requires coordinated player interaction, enforcing synchronization and shared decision-making before advancing the mission.

To simulate perceptual stress and uncertainty, we incorporate visibility constraints. Object detectability is limited by spatial distance and modulated by player health, reflecting reduced perceptual capability under strain. As health declines, the probability of perceiving nearby objects decreases, introducing uncertainty into search and decision-making. Elevator system and visibility constraints provide structured task control while modeling realistic perceptual challenges.

6 MULTI-USER COLLABORATION

To enable systematic evaluation of cognitive attacks in realistic operational settings, we extend the CAESAR testbed to support multi-user MR collaboration. Search and rescue tasks depend on shared situational awareness and coordinated action, which introduce additional synchronization and shared-state dependencies beyond single-user interaction. Our multi-user design provides shared spatial alignment and real-time synchronization, enabling participants to operate within a consistent environment while exposing collaborative attack surfaces such as desynchronization and shared-perception manipulation.

6.1 Colocation System and Spatial Alignment

A consistent spatial reference frame is essential for collaborative MR tasks, as misalignment can disrupt shared perception and coordination. To ensure spatial consistency, users first reconstruct their physical environment and join a network session. We then employ shared spatial anchors [27] to align coordinate systems across devices, transforming each local frame into a unified global reference. Once aligned, all participants perceive virtual objects at identical spatial locations relative to the physical environment. Visual feedback confirms successful alignment, ensuring that spawning, animation, and interaction remain

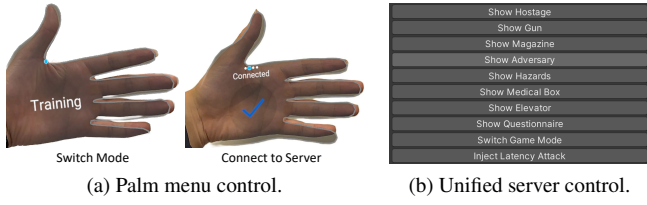


Figure 5: Meta control mechanisms.

consistent across users. This colocation mechanism forms the foundation for reliable shared situational awareness and enables controlled study of collaborative cognitive attack surfaces.

6.2 Real-Time Synchronization

To preserve shared situational awareness, the platform synchronizes both virtual objects and player-related states across all participants. Interactive objects, including hostages, adversaries, hazards, and weapons, are instantiated and updated consistently so that all users observe identical environmental dynamics. Concurrent interactions are resolved through authority-controlled state management to prevent divergence across clients. In addition to environmental elements, the system continuously synchronizes player-specific information, including health, six-degree-of-freedom (6DoF) head pose, and mission-level metrics (e.g., adversaries defeated and hostages rescued). These updates are reflected in real time on each participant’s interface, ensuring a consistent perception of team status. This synchronization architecture maintains coherence in the shared environment and is essential for studying collaborative cognitive attack surfaces, where inconsistencies in state propagation, authority control, or perception could directly influence coordination, trust, and decision-making.

7 META CONTROL AND DATA COLLECTION FRAMEWORK

7.1 Meta Control

To facilitate experimental user studies, the CAESAR testbed integrates meta-level control mechanisms that enable developer testing and centralized coordination. At the local level, participants interact with a palm-menu interface, while at the global level, a unified server control manages training flow and cognitive attack.

Palm Menu Control. As Figure 5a shows, users can summon a palm-mounted menu to switch between modes, manage server connections, and configure study parameters. The menu provides a training mode that familiarizes participants with virtual objects, interactions, and task mechanics, as well as experiment modes that instantiate predefined scenario configurations (e.g., object types and spatial distributions). Key parameters (e.g., the number of hostages and adversaries) can be adjusted at runtime, enabling rapid iteration over experimental conditions without rebuilding the scene.

Unified Server Control. We designed the unified server control to facilitate training and manage cognitive attacks during experiment modes, as Figure 5b shows. In training mode, the server selectively enables virtual objects and interactions to standardize participant onboarding. In experiment modes, the server manages mode transitions and centrally coordinates cognitive attack injection, ensuring synchronized delivery across participants and consistent timing across trials.

7.2 Data Collection Pipeline

The CAESAR testbed integrates a synchronized, multimodal data collection pipeline that records system- and user-level metrics in real time, including cognitive attack events, behavioral data, user activity events, and subjective measures. In addition, the pipeline captures first-person video and in-game audio communication, providing contextual grounding for quantitative analysis. The server stores all data for subsequent processing, providing a synchronized multimodal dataset that supports fine-grained post-hoc analysis of user resilience, decision-making, and collaboration behaviors under adversarial MR conditions. The data collection pipeline is shown in Figure 6.

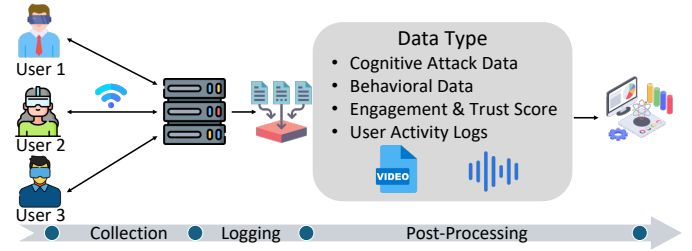


Figure 6: Data collection pipeline. Multimodal data are captured and logged on a centralized server for post-hoc analysis.

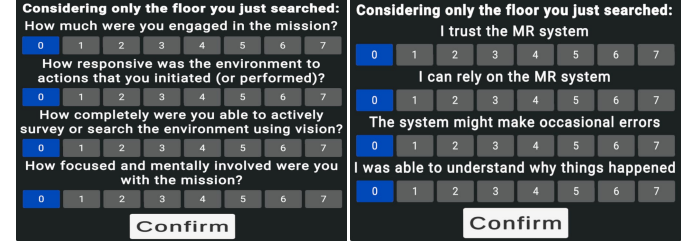


Figure 7: Pop-up questionnaire mechanism. At predetermined intervals, the system displays a pop-up to collect 7-point Likert scale ratings of user engagement and trust.

Real-Time Data Collection. During experiment modes, the platform continuously streams system-level and user-level data to the centralized server. Cognitive attacks are injected at controlled intervals, and each perturbation event is explicitly time-stamped and labeled by type and intensity. This structured annotation enables direct alignment between attack exposure and subsequent behavioral or subjective responses, supporting systematic evaluation of attack impact.

In parallel, the system captures rich behavioral data, including head and eye tracking. Specifically, the system continuously records the position and rotation of the head and eyes, gaze direction, and 2D focal points, along with derived metrics such as gaze targets and information about visible virtual objects (e.g., object count, name, and pixel coordinates within the user’s field of view). These data provide fine-grained insights into users’ attention allocation and cognitive load.

To collect subjective measures such as engagement and trust, synchronized pop-up questionnaires (shown in Figure 7) appear in participants’ heads-up display (HUD) at fixed intervals. Questionnaire presentation is synchronized across all participants to ensure temporally aligned reporting and consistent exposure windows. Responses use a 7-point Likert scale, with a default value of 0 indicating an invalid response to prevent bypassing.

The system also logs user interaction events in real time. Actions such as defeating adversaries, rescuing hostages, or using medical boxes are time-stamped and annotated with structured event labels, enabling precise reconstruction of behavioral trajectories and decision sequences during post-hoc analysis.

Data Logging and Post-Processing. All participant data streams are collected through a centralized Python-based server prior to local storage. Time-stamps are aligned across behavioral logs, cognitive attack events, questionnaire responses, and first-person video, enabling cross-modal synchronization. A post-processing pipeline integrates these heterogeneous data sources into a unified dataset structured for quantitative and qualitative analysis. The modular schema allows additional attack types or behavioral signals to be incorporated without restructuring the logging framework, supporting extensibility for future adversarial MR studies.

8 USER STUDY: LATENCY-BASED COGNITIVE ATTACK

While the CAESAR testbed supports a broad range of cognitive attacks across multiple categories, conducting comprehensive human-subject evaluations for all attack types within a single paper is infeasible. Instead, we focus our user study on latency-based display perturbation, specifically the frame rate reduction and action-to-display latency mech-

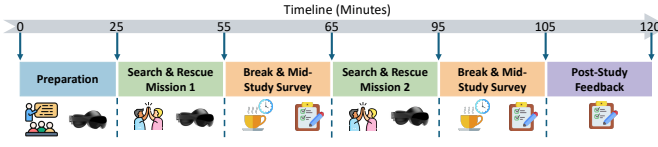


Figure 8: User study timeline.

anisms described in Section 3. This attack serves as a representative evaluation case within our framework because it directly interferes with the perception-action feedback loop essential for situational awareness and coordination in MR environments.

We conducted a controlled human-subject study with 22 participants to systematically examine the impact of latency manipulation while demonstrating the framework’s capability to support controlled human-subject evaluation. The study integrates objective task-performance metrics, fine-grained behavioral signals, and subjective measures of engagement and trust. This multi-level measurement design enables a comprehensive analysis of how display latency affects perception, decision-making, and collaborative performance in MR-based search-and-rescue tasks. The same experimental protocol and data-collection pipeline can be readily extended to evaluate other cognitive attack types supported by CAESAR in future work.

8.1 Research Questions for the User Study

The user study is designed to investigate how the latency-based cognitive attack affects user performance, behavior, and subjective experience during collaborative MR search-and-rescue tasks. Specifically, we address the following research questions:

- **RQ1:** *How does injected latency affect task performance in collaborative MR search-and-rescue scenarios?*
- **RQ2:** *How do users adapt their behavior under increasing latency during collaborative MR tasks?*
- **RQ3:** *How does latency-based cognitive attack affect users’ subjective experience, including engagement and trust?*

These research questions aim to provide a comprehensive understanding of the cognitive and behavioral impact of the latency-based attack in MR training environments, while demonstrating the capability of the CAESAR testbed to support systematic, human-centered evaluation of cognitive attacks.

8.2 Participants

We recruited 22 participants for the user study (6 female, 15 male, 1 non-binary), ranging in age from 19 to 32 years, with an average age of 24.4 and a standard deviation of 3.8. Participants had normal or corrected-to-normal vision and no self-reported motor impairments that would interfere with MR interaction. All participants provided informed consent prior to participation, and the study protocol was approved by the local Institutional Review Board (IRB).

Participants completed the study as part of two-person teams in a collaborative MR search-and-rescue session, labeled as Participant A and Participant B for analysis. During the task, teammates were required to coordinate navigation, adversary engagement, and hostage rescue under varying latency conditions. Each participant followed the same structured protocol consisting of onboarding, a training phase, and experimental trials conducted under controlled conditions to reduce variability due to learning effects. Participants received \$40 cash compensation for their time.

8.3 Experimental Design

Study Overview. The user study followed a structured experimental design lasting approximately 120 minutes per participant. The session consisted of a preparation and training phase, two collaborative search-and-rescue missions, two intervening breaks, and a post-study feedback phase, as shown in Figure 8. Participants were additionally invited to complete an optional online follow-up survey the following day to report any perceived after-effects. This design balances realistic task dynamics with controlled latency manipulation while mitigating fatigue in a long-duration MR study.

Preparation and Training. Participants first completed a standardized preparation and training phase designed to familiarize them with the MR hardware, interaction techniques, and task mechanics, thereby reducing learning effects during the formal study. Training covered core interactions required during the study, including hand-based interaction, adversary engagement, hostage rescue, hazard avoidance, and use of the virtual elevator system. Participants also practiced responding to pop-up engagement and trust questionnaires presented within the MR environment. The training protocol followed a fixed script and was identical across sessions to ensure procedural consistency.

Search-and-Rescue Missions. Participants completed two collaborative search-and-rescue missions, each lasting approximately 25 minutes, within a virtual five-floor building. During each mission, teams searched for and rescued hostages while managing hazards and adversaries under time pressure. Each virtual floor lasted approximately four minutes. Upon completion of each floor, participants responded to synchronized pop-up questionnaires assessing engagement and trust, resulting in five prompts per mission.

To ensure fair comparisons across latency conditions and participants, a predefined mission configuration was assigned to each floor. While the spatial locations of virtual objects varied across floors to reduce memorization effects, the number and types of virtual objects remained consistent to preserve comparable task composition. Each floor corresponded to a distinct latency condition. Latency-based cognitive attack was applied by perturbing display responsiveness, including controlled action-to-display delays and reduced frame rates, enabling systematic analysis of latency effects on task performance, behavior, and subjective experience.

Breaks and Feedback. Two 10-minute breaks were scheduled between missions, during which participants removed the MR headset and completed questionnaires. These breaks were intended to reduce physical and cognitive fatigue and mitigate potential simulator sickness. Following the second break, participants completed demographic questionnaires and provided general feedback on their experience.

Follow-on Survey. The day after the study, participants were invited to complete a brief optional online survey evaluating any after-effects within 12 hours following participation. The survey consisted of a binarized form of the Simulator Sickness Questionnaire (SSQ) [22].

8.4 Measures

To evaluate the impact of latency-based cognitive attack on collaborative MR performance and user experience, we analyze objective task performance metrics, behavioral indicators, and subjective measures collected during the study.

Objective Task Performance. Objective task performance was evaluated for each search-and-rescue mission, including: (1) *Mission completion outcomes*, including the team/floor-level hostage rescue rate and the number of adversaries defeated; (2) *Task efficiency*, measured by the team/floor-level time to first hostage rescue, and the exploratory rescue-event-level per-rescue completion time; (3) *Safety-efficiency trade-offs*, quantified through hostage health loss events relative to rescue timing. These metrics reflect how latency perturbations affect participants’ ability to complete mission objectives under time pressure.

Behavioral Measures. Behavioral indicators were derived from in-mission activity, including: (1) *Interaction behavior*, such as the frequency of weapon-use actions; (2) *Exposure and state-related behavior*, assessed through hazard exposure events. These behavioral measures characterize how participants adapt their interaction strategies under varying levels of latency.

Subjective Measures. Participants’ subjective experience was evaluated using self-reported engagement and trust ratings collected through repeated in-task questionnaires and post-study surveys: *Engagement and trust*, measured as participant/floor-level repeated measures through pop-up questionnaires after each mission floor, yielding five responses per mission. These responses capture both momentary fluctuations and cumulative effects of latency-based cognitive attacks on user experience.

Table 1: Team-Level Rescue Time per Hostage by Latency Level

Latency (ms)	Count	Mean (s)	SD (s)	Median (s)
0	11	56.152	41.276	33.034
9	21	40.113	28.192	29.999
17	15	61.760	52.994	37.952
25	28	34.571	19.572	30.007
62	30	36.957	20.283	30.025

8.5 Results

We report results from the 22-participant user study examining the effects of the latency-based cognitive attack on collaborative MR performance, behavior, and subjective experience. Although two missions were scheduled, some teams did not complete the second mission within the session time. Therefore, all reported analyses are based on data from the first mission to ensure consistency across participants. In this mission, latency levels were assigned to floors in a fixed sequence (0, 17, 62, 9, and 25 ms from floors 1 to 5). Results are organized according to task performance, behavioral adaptation, and subjective measures, corresponding to the research questions outlined in Section 8.1.

Task Performance. Figure 9 summarizes team-level task performance under different injected latency levels. Overall, the results suggest that participants adapted their task strategies under higher latency, allowing teams to maintain high rescue rates despite increased perceptual and interaction delays. However, latency continued to influence task dynamics and rescue quality, revealing a trade-off between successful task completion and the quality of task outcomes.

Rescue Rate. As shown in Figure 9 (Left), rescue rate differs significantly across latency conditions (Friedman test: $\chi^2(4) = 30.07$, $p < 0.001$, Kendall’s $W = 0.68$; generalized estimating equations (GEE) Wald test: $\chi^2(4) = 45.28$, $p < 0.001$). Interestingly, the highest rescue rates are observed under the strongest latency perturbations (25 ms and 62 ms), a counter-intuitive result given that increased latency would generally be expected to impair task performance. As discussed in the following analysis of time-to-first-rescue, participants adapted their task strategies under higher latency by initiating rescues earlier in the mission, which likely compensated for increased perceptual and interaction delays and contributed to the observed rescue performance.

Adversaries Defeated. Although the total number of adversaries defeated is evaluated as a team/floor-level metric, only one weapon is available per floor. Therefore, in addition to the team/floor-level metric, Figure 9 (Center left) presents the distribution of adversaries defeated at the participant level to illustrate how combat responsibility is allocated between teammates under different latency conditions. The median number of adversaries defeated remains relatively stable across conditions, with no clear monotonic relationship between latency and adversaries defeated. However, higher latency levels (particularly 25 ms) exhibit increased dispersion for certain participants, indicating greater variability in individual combat engagement. This pattern suggests that latency does not uniformly reduce adversary defeat rates, but may influence how participants allocate effort between combat and rescue activities.

Rescue Time. As shown in Figure 9 (Center right), the time to first rescue differs significantly across latency conditions (Friedman test: $\chi^2(4) = 17.00$, $p = 0.0019$, Kendall’s $W = 0.53$). A linear mixed-effects model also identified significantly shorter times to first rescue under the 9 ms ($p = 0.001$), 25 ms ($p < 0.001$), and 62 ms ($p < 0.001$) latency conditions relative to the baseline. Interestingly, teams generally initiated rescues earlier under higher latency conditions rather than exhibiting progressively slower performance. These findings help explain the rescue-rate results discussed above. By prioritizing rescue objectives earlier in the mission, participants were able to compensate for perceptual and interaction delays, resulting in consistently high, and in some cases higher, rescue rates despite stronger latency perturbations. A similar trend is observed in the exploratory team-level per-rescue completion time analysis (see Table 1), where the 25 ms and 62 ms conditions exhibit lower average completion times than several lower-latency conditions. These results suggest that latency primarily influenced task strategy rather than simply degrading task execution.

Rescue Time and Hostage Health Trade-off. Although participants adapted their rescue strategies under higher latency conditions, longer rescue durations remained associated with reduced rescue quality. As shown in Figure 9 (Right), rescue duration is negatively correlated with hostage health at the time of rescue ($r = -0.28$, $p \approx 0.0042$), highlighting a time-critical trade-off inherent to the task. This relationship persists across latency conditions, suggesting that while participants adapt to latency and maintain rescue rate, longer rescue durations continued to be associated with lower remaining hostage health. Overall, these findings indicate that maintaining rescue rate does not eliminate the quality cost associated with longer rescue durations.

Behavioral Adaptation. Figure 10 shows how participant behavior changes under different latency levels, focusing on weapon interaction patterns and exposure to environmental hazards. Unlike task performance metrics, these behavioral measures reveal behavioral adaptations in response to latency-based cognitive attacks.

Weapon Usage Patterns. As shown in Figure 10a, weapon actions exhibit a non-linear relationship with latency. Moderate latency levels (e.g., 17 and 25 ms) are associated with an increase in both shots fired and reload actions, suggesting increased action frequency under moderate latency. This behavior may reflect perceived uncertainty in action timing and feedback, leading users to fire more frequently or reload proactively to maintain combat readiness. At the highest latency level (62 ms), however, total weapon actions decrease, consistent with a more conservative pattern of combat engagement. These trends suggest that latency does not merely slow interactions but may also influence how participants regulate engagement intensity during combat.

Hazard Exposure Time. Figure 10b shows the distribution of hazard exposure time across latency conditions, capturing the duration participants spend in fire or poison states. The no-latency condition exhibits the greatest spread in exposure duration, whereas moderate latency levels (17 ms and 25 ms) show lower median exposure times and comparatively tighter distributions. The highest latency level (62 ms) demonstrates moderate variability but does not exceed that observed in the no-latency condition. This pattern suggests more consistent hazard avoidance at moderate latency levels, while larger delays appear to affect responsiveness during hazard avoidance. These findings indicate that latency may influence the consistency of hazard avoidance behavior rather than producing systematic degradation.

Subjective Experience. Figure 11 presents participants’ subjective responses under different latency conditions, focusing on perceived engagement and trust in the system. In contrast to task performance and behavioral measures, these subjective ratings capture participants’ perceived interaction quality and system reliability. Despite statistically significant differences across latency conditions, engagement and trust remain generally stable, suggesting a limited impact on participants’ subjective experience.

Engagement. As shown in Figure 11a, engagement differs significantly across latency conditions (Friedman test: $\chi^2(4) = 29.37$, $p < 0.001$, Kendall’s $W = 0.33$). A linear mixed-effects model similarly identified significantly higher engagement scores under all latency conditions relative to the baseline (17 ms: $p = 0.013$; 9 ms, 25 ms, and 62 ms: all $p < 0.001$). Interestingly, participants reported slightly higher engagement under latency than under the no-latency condition, despite the added perceptual and interaction challenges. Although these differences are statistically significant, average engagement scores remained consistently high across all conditions. These findings suggest that the evaluated latency-based cognitive attacks did not reduce perceived engagement and may have slightly increased task involvement.

Trust. As shown in Figure 11b, trust differs significantly across latency conditions (Friedman test: $\chi^2(4) = 11.51$, $p = 0.021$, Kendall’s $W = 0.13$). A linear mixed-effects model similarly identified slightly higher trust scores under the 9 ms ($p = 0.004$), 25 ms ($p < 0.001$), and 62 ms ($p = 0.002$) latency conditions relative to the baseline, while the 17 ms condition did not differ significantly ($p = 0.146$). Despite these statistically significant differences, trust scores remained relatively stable across latency conditions, with similar median values and substantial overlap in their distributions. These findings suggest that the evaluated latency-based cognitive attacks did not substantially reduce

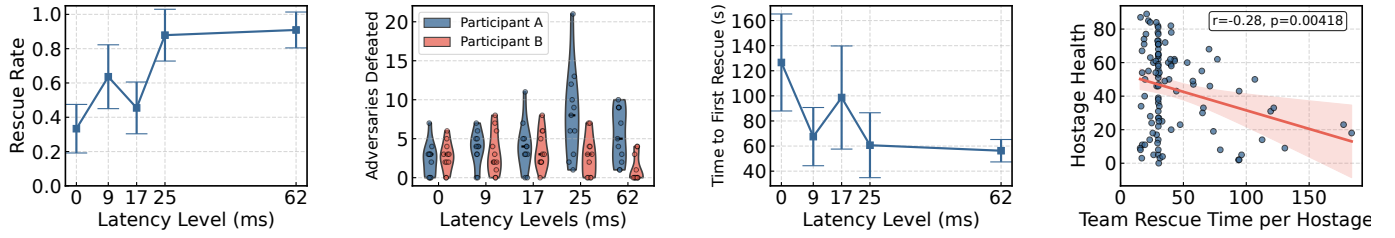


Figure 9: Task performance under latency-based cognitive attack. **(Left)**: Team rescue rate as a function of injected latency, showing mean and 95% confidence intervals. **(Center left)**: Distribution of adversaries defeated per participant under different latency levels. **(Center right)**: Time to first successful rescue across latency conditions. **(Right)**: Relationship between team-level rescue duration and hostage health at rescue, with linear regression and 95% confidence band.

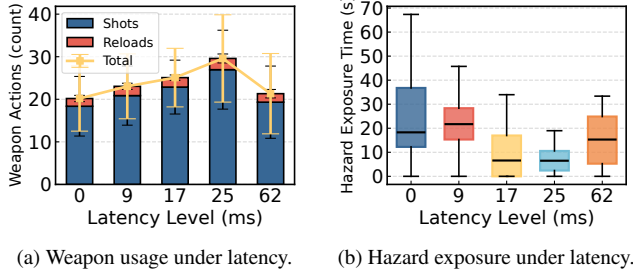


Figure 10: Behavioral adaptation under latency-based cognitive attack. **(Left)**: Changes in weapon interaction patterns, including shots fired and reload actions. **(Right)**: Distribution of hazard exposure time (fire and poison) across latency conditions.

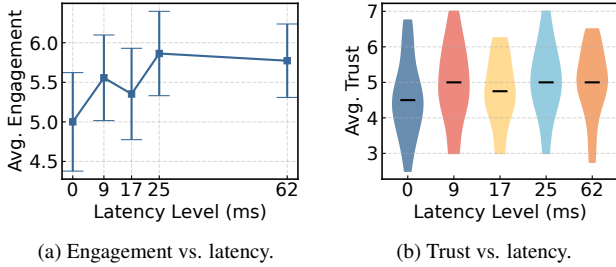


Figure 11: Subjective experience under latency-based cognitive attack. **(Left)**: Mean engagement scores across latency conditions. **(Right)**: Distribution of trust scores across latency conditions.

perceived trust, and the observed differences were relatively modest.

Self-reported After-effects. 10 out of 22 participants completed the optional binarized SSQ the day after the study. Only one participant reported a headache within 12 hours of participation. Given the limited responses, further investigation is needed to determine whether CAESAR may contribute to post-session symptoms.

9 DISCUSSION AND FUTURE WORK

This section discusses current limitations of the framework and outlines directions for future research and system enhancement.

Experimental Design Considerations. The current user study employs a fixed ordering of latency conditions to ensure identical task configurations across participants. While this design introduces the possibility of order effects, the observed results are unlikely to be explained solely by participant familiarization. Repeated-measures analyses identified significant effects of latency on both rescue rate and time to first rescue, and the behavioral analyses suggest that participants adapted their task strategies under higher latency conditions by initiating rescues earlier. Randomized condition ordering would further reduce potential order effects and will be considered in future studies, although such a design would require substantially more participants to maintain balanced experimental conditions and statistical power.

Attack Detection and Mitigation. Our proposed testbed provides a unified platform for injecting multiple cognitive attacks within a controlled MR environment, enabling system-level evaluation under adversarial conditions. However, the current implementation focuses

on systematic attack deployment and evaluation rather than real-time detection or mitigation. Although detection modules are not yet integrated, the modular architecture of the testbed is explicitly designed to support such extensions. Future work will incorporate detection and response mechanisms to enable closed-loop evaluation of adversarial robustness. For example, prior works [41–43] have explored visual-language model (VLM)-based approaches for detecting adversarial or misleading AR content, which could be integrated as detection components and systematically evaluated within our platform.

Scalability and Generalization. Although we use collaborative search-and-rescue as a representative application scenario, the proposed framework is not limited to this domain. Search-and-rescue was selected because it integrates multiple core components of collaborative MR applications, including target search and localization, navigation under hazards, combat interaction, multi-user collaboration, synchronization, decision-making, and behavioral adaptation. Furthermore, the framework adopts a modular architecture that separates attack injection, synchronization and colocation, experiment orchestration, data logging, and task content, enabling adaptation to other collaborative MR applications such as industrial target search, fire evacuation training, and tactical shooting training. In future work, generative AI techniques [46] could be leveraged to accelerate scenario construction and virtual content generation, reducing manual design effort and enabling more diverse task configurations. While the current user study focuses on latency-based perturbations as a representative case study, future work will evaluate the remaining cognitive attack classes implemented in the framework to systematically investigate how diverse adversarial conditions influence users’ perception, behavior, coordination, and trust in collaborative MR settings.

Multimodal Behavioral Analysis. The testbed supports synchronized collection of multimodal data, enabling comprehensive analysis of user behavior during collaborative MR tasks. This extensible data collection pipeline facilitates quantitative evaluation of both task performance and user experience under adversarial or degraded conditions rather than in isolated settings. Furthermore, these multimodal data can support future research on how cognitive attacks influence visual attention, situational awareness, and decision-making in collaborative MR environments.

10 CONCLUSION

We presented the CAESAR testbed, an end-to-end multi-user mixed reality framework for systematic evaluation of cognitive attacks in collaborative search-and-rescue training. The framework adopts a modular architecture that supports synchronized collaboration, integrates four representative attack classes, and incorporates a multimodal data collection pipeline for analyzing user behavior and performance under adversarial conditions. Through a controlled user study of the latency-based attack, we demonstrated measurable effects on task performance and user behavior while characterizing users’ subjective experience under adversarial conditions. Looking forward, CAESAR provides a reproducible foundation for collaborative MR security research by enabling controlled attack injection, synchronized behavioral measurement, and systematic evaluation of adversarial effects.

ACKNOWLEDGMENTS

This work has been supported by DARPA grant HR0011-2420366.

REFERENCES

- [1] A. Augello, G. Caggianese, L. Gallo, and P. Neroni. Design strategies to enhance awareness in MR collaborative systems. In *International Conference on Extended Reality*, pp. 424–436. Springer, 2023. doi: 10.1007/978-3-031-43401-3_28 2
- [2] E. Azimi, A. Winkler, E. Tucker, L. Qian, J. Doswell, N. Navab et al. Can mixed-reality improve the training of medical procedures? In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4065–4068. IEEE, 2018. doi: 10.1109/EMBC.2018.8513387 1
- [3] W. Büschel, A. Lehmann, and R. Dachselt. MIRIA: A mixed reality toolkit for the in-situ visualization and analysis of spatio-temporal interaction data. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–15, 2021. doi: 10.1145/3411764.3445651 2
- [4] J. Chen, M. Zhu, and B. Li. PRISM-XR: Empowering privacy-aware XR collaboration with multimodal large language models. *arXiv preprint arXiv:2602.10154*, 2026. doi: 10.48550/arXiv.2602.10154 2
- [5] K. Cheng, A. Bhattacharya, M. Lin, J. Lee, A. Kumar, J. F. Tian et al. When the user is inside the user interface: An empirical study of UI security properties in augmented reality. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 2707–2723. USENIX Association, Philadelphia, PA, Aug. 2024. 2, 3
- [6] K. Cheng, J. F. Tian, T. Kohno, and F. Roesner. Exploring user reactions and mental models towards perceptual manipulation attacks in mixed reality. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 911–928. USENIX Association, Anaheim, CA, Aug. 2023. 2, 3
- [7] Z. Dong, J. Chen, and B. Li. Demo: Collaborative mixed-reality-based firefighter training. In *2023 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 1–2, 2023. doi: 10.1109/INFOCOMWKSHPS57453.2023.10226009 1, 2
- [8] D.-T. Duong-Le, T.-T. Tran, H.-N. Do, D.-N. Ly, M.-T. Tran, and K.-D. Le. MRClassroom: A mixed-reality interface for improving remote students’ presence in hybrid classrooms. In *International Symposium on Information and Communication Technology*, pp. 338–352. Springer, 2024. doi: 10.1007/978-981-96-4288-5_27 1
- [9] M. El-Hajj. Cybersecurity and privacy challenges in extended reality: Threats, solutions, and risk mitigation strategies. In *Virtual Worlds*, vol. 4, p. 1. MDPI, 2024. doi: 10.3390/virtualworlds4010001 3
- [10] B. Ens, J. Lanir, A. Tang, S. Bateman, G. Lee, T. Piumsomboon et al. Revisiting collaboration through mixed reality: The evolution of groupware. *International Journal of Human-Computer Studies*, 131:81–98, 2019. doi: 10.1016/j.ijhcs.2019.05.011 2
- [11] Z. Ersozlu, S. Ledger, A. Ersozlu, F. Mayne, and H. Wildy. Mixed-reality learning environments in teacher education: An analysis of TeachLivE™ research. *Sage Open*, 11(3):21582440211032155, 2021. doi: 10.1177/21582440211032155 1
- [12] J. A. Frank and V. Kapila. Mixed-reality learning environments: Integrating mobile interfaces with laboratory test-beds. *Computers & Education*, 110:88–104, 2017. doi: 10.1016/j.compedu.2017.02.009 1
- [13] G. M. Garrido, V. Nair, and D. Song. SoK: Data privacy in virtual reality. *arXiv preprint arXiv:2301.05940*, 2023. doi: 10.56553/popets-2024-0003 3
- [14] M. Gonzalez-Franco, R. Pizarro, J. Cermeron, K. Li, J. Thorn, W. Hutabarat et al. Immersive mixed reality for manufacturing training. *Frontiers in Robotics and AI*, 4:3, 2017. doi: 10.3389/frobt.2017.00003 1
- [15] M. Gottsacker, N. Numan, A. Steed, G. Bruder, G. F. Welch, and S. Feiner. Decoupled hands: An approach for aligning perspectives in collaborative mixed reality. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–7, 2025. doi: 10.1145/3706599.3720219 2
- [16] J. E. S. Grønbaek, J. Sánchez Esquivel, G. Leiva, E. Velloso, H. Gellersen, and K. Pfeuffer. Blended whiteboard: Physicality and reconfigurability in remote mixed reality collaboration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2024. doi: 10.1145/3613904.3642293 2
- [17] J. Happa, M. Glencross, and A. Steed. Cyber security threats and challenges in collaborative mixed-reality. *Frontiers in ICT*, 6:5, 2019. doi: 10.3389/fict.2019.00005 3
- [18] B. Hu, Y. Zhang, S. Hao, and Y. Tao. MOFA: Exploring asymmetric mixed reality design strategy for co-located multiplayer between handheld and head-mounted augmented reality. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–4, 2023. doi: 10.1145/3544549.3583935 2
- [19] M. Huzaifa, R. Desai, S. Grayson, X. Jiang, Y. Jing, J. Lee et al. Illuxr: Enabling end-to-end extended reality research. In *2021 IEEE International Symposium on Workload Characterization (IISWC)*, pp. 24–38, 2021. doi: 10.1109/IISWC53511.2021.00014 3
- [20] P. Jansen, F. Fischbach, J. Gugenheimer, E. Stemasov, J. Frommel, and E. Rukzio. ShARe: Enabling co-located asymmetric multi-user interaction for augmented reality head-mounted displays. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pp. 459–471, 2020. doi: 10.1145/3379337.3415843 2
- [21] X. Ji, H. Liu, Z. Wang, Y. Li, Y. Tian, and M. Huo. Agent-based MR simulation as a tool for combat search and rescue command and control. In *Proceedings of the 2019 3rd International Conference on Artificial Intelligence and Virtual Reality, AIVR 2019*, pp. 1–7. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3348488.3348489 2
- [22] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology*, 3(3):203–220, 1993. doi: 10.1207/s15327108ijap0303_3 7
- [23] H. Li, I. Oguntola, D. Hughes, M. Lewis, and K. Sycara. Theory of mind modeling in search and rescue teams. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 483–489, 2022. doi: 10.1109/RO-MAN53752.2022.9900572 1, 2
- [24] M. Lorenz, M. Kaluschke, A. Melzer, N. Pillen, M. Sanrow, A. Hoffmann et al. HIPS—A surgical virtual reality training system for total hip arthroplasty (THA) with realistic force feedback. *IEEE Transactions on Visualization and Computer Graphics*, 2025. doi: 10.1109/TVCG.2025.3549896 1, 2
- [25] Meta. Mixed Reality Utility Kit | Meta Horizon OS Developers. <https://developers.meta.com/horizon/documentation/unity/unity-mr-utility-kit-overview/>. 4
- [26] Meta. Passthrough | Meta Horizon OS Developers. <https://developers.meta.com/horizon/documentation/unity/unity-passthrough/>. 3
- [27] Meta. Shared Spatial Anchors | Meta Horizon OS Developers. <https://developers.meta.com/horizon/documentation/unity/unity-shared-spatial-anchors/>. 4, 5
- [28] R. Miller, N. K. Banerjee, and S. Banerjee. Temporal effects in motion behavior for virtual reality (VR) biometrics. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 563–572, 2022. doi: 10.1109/VR51125.2022.00076 4
- [29] M. P. Oppelt, A. Foltyn, J. Deuschel, N. R. Lang, N. Holzer, B. M. Eskofier et al. ADABase: a multimodal dataset for cognitive load estimation. *Sensors*, 23(1):340, 2022. doi: 10.3390/s23010340 2
- [30] N. Pellas, I. Kazanidis, and G. Palaigeorgiou. A systematic literature review of mixed reality environments in K-12 education. *Education and Information Technologies*, 25(4):2481–2520, 2020. doi: 10.1007/s10639-019-10076-4 1
- [31] Photon Engine. Photon Fusion. <https://doc.photonengine.com/zh-cn/fusion/current/fusion-2-intro>. 4
- [32] I. Safranoglou, A. Stavroulakis, M. Ebel, J. Pottebaum, G. Lamprinakos, D. Dimelli et al. Augmented reality for real-time decision-making in flood emergencies. In *2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 110–116. IEEE, 2024. doi: 10.1109/ISMAR-Adjunct64951.2024.00032 1
- [33] M. Sajid, S. I. M. S. Bukhari, B. Ji, and B. David-John. “Just stop doing everything for now!”: Understanding security attacks in remote collaborative mixed reality. In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 623–633. IEEE, 2025. doi: 10.1109/VR59515.2025.00085 2, 3
- [34] L. Sidenmark, T. Zhang, L. Al Lababidi, J. Li, and T. Grossman. Desk2Desk: Optimization-based mixed reality workspace integration for remote side-by-side collaboration. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–15, 2024. doi: 10.1145/3654777.3676339 2
- [35] A. Teymourian, A. M. Webb, T. Gharaibeh, A. Ghildiyal, and I. Baggili. SoK: Come together – unifying security, information theory, and cognition for a mixed reality deception attack ontology & analysis framework. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 1475–1492. USENIX Association, Seattle, WA, Aug. 2025. 2, 3
- [36] U.S. Army. Training circular 3-37.51: Urban search and rescue. Technical

report, Department of the Army, Washington D.C., USA, 2020. [1](#), [2](#)

- [37] M. Wilchek, K. Luther, and F. A. Batarseh. Ajna: A wearable shared perception system for extreme sensemaking. *ACM Trans. Interact. Intell. Syst.*, 15(1), art. no. 1, 29 pp., Jan. 2025. doi: 10.1145/3690829 [2](#)
- [38] M. Wilchek, L. Wang, S. Dickinson, E. Feuerbacher, K. Luther, and F. A. Batarseh. KHAIT: K-9 handler artificial intelligence teaming for collaborative sensemaking. In *Proceedings of the 30th International Conference on Intelligent User Interfaces, IUI '25*, pp. 925–937. Association for Computing Machinery, 2025. doi: 10.1145/3708359.3712107 [2](#), [3](#)
- [39] M. M. Willett and M. Demir. Understanding the impact of team cognitive load and advice compliance in urban search and rescue task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 67(1):2484–2489, 2023. doi: 10.1177/21695067231192293 [1](#), [2](#)
- [40] E. Wilson, A. Ibragimov, M. J. Proulx, S. D. Tetali, K. Butler, and E. Jain. Privacy-preserving gaze data streaming in immersive interactive virtual reality: Robustness and user experience. *IEEE Transactions on Visualization and Computer Graphics*, 30(5):2257–2268, 2024. doi: 10.1109/TVCG.2024.3372032 [4](#)
- [41] Y. Xiu and M. Gorlatova. Detecting visual information manipulation attacks in augmented reality: a multimodal semantic reasoning approach. *IEEE Transactions on Visualization and Computer Graphics*, 2025. doi: 10.1109/TVCG.2025.3616842 [2](#), [9](#)
- [42] Y. Xiu, T. Scargill, and M. Gorlatova. LOBSTAR: Language model-based obstruction detection for augmented reality. In *2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 335–336. IEEE, 2024. doi: 10.1109/ISMAR-Adjunct64951.2024.00078 [2](#), [9](#)
- [43] Y. Xiu, T. Scargill, and M. Gorlatova. ViDDAR: Vision language model-based task-detrimental content detection for augmented reality. *IEEE transactions on visualization and computer graphics*, 2025. doi: 10.1109/TVCG.2025.3549147 [2](#), [9](#)
- [44] X. Yao, J. Chen, T. He, J. Yang, and B. Li. A scalable mixed reality platform for remote collaborative LEGO design. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 1–2. IEEE, 2022. doi: 10.1109/INFOCOMWKSHPS54753.2022.9798010 [2](#)
- [45] J. Zhao, P. LaFemina, J. Carr, P. Sajjadi, J. O. Wallgrün, and A. Klippel. Learning in the field: Comparison of desktop, immersive virtual reality, and actual field trips for place-based STEM education. In *2020 IEEE conference on virtual reality and 3D user interfaces (VR)*, pp. 893–902. IEEE, 2020. doi: 10.1109/VR46266.2020.00012 [1](#), [2](#)
- [46] M. Zhu, J. Chen, and B. Li. When generative artificial intelligence meets extended reality: Enabling scalable and natural interactions. *IEEE Internet Computing*, 29(6):15–24, 2025. doi: 10.1109/MIC.2025.3619462 [9](#)