Believe It. Engage It.: A Mixed Reality Dataset of Engagement, Cognitive Load, Physical Load, Cybersickness, Task Performance Under Cognitive Attack

Nasim Ahmed* Kennesaw State University
Md Mahedi Hassan† Kennesaw State University
Peng Wu‡ Northeastern University
Kaiming Huang§ The Pennsylvania State University
Mingyu Zhu¶ The Pennsylvania State University
JoAnn Archer‖ Design Interactive, Inc.
John Kwasinski** Design Interactive, Inc.
Peyton S. Bailey†† Design Interactive, Inc.
Jessyca L. Derby‡‡ Design Interactive, Inc.
Tian Lan§§ George Washington University
Bin Li¶¶ The Pennsylvania State University
Gang Tan‖ The Pennsylvania State University
Mahdi Imani*** Northeastern University
Rifatul Islam††† Kennesaw State University

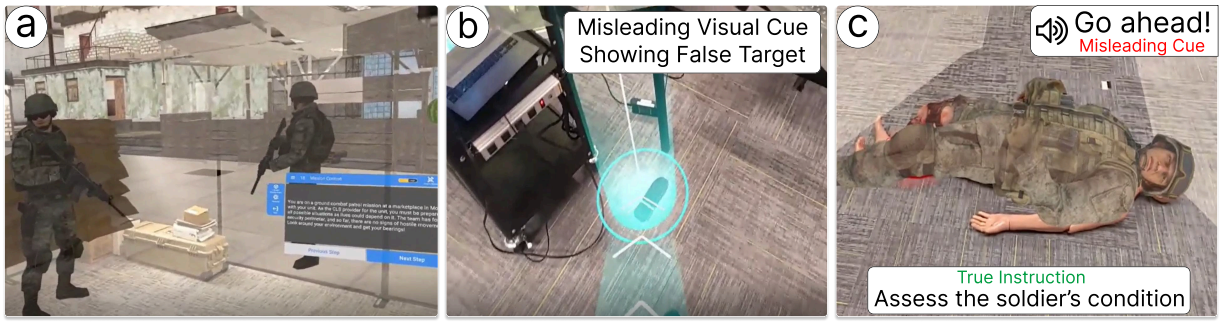

Figure 1: **Mixed Reality training environment and deceptive guidance (cognitive attack) conditions.** (a) Baseline Tactical Combat Casualty Care (TCCC) Lifesaver training scenario where the system provides correct instructions for applying a tourniquet. (b) Visual deception condition where the interface highlights an incorrect location on the ground as the target. The highlighted object is not the correct item needed for the procedure, which can mislead the user during the task. (c) Auditory deception condition where the system plays the voice command "Go ahead" even though the correct instruction is to first assess the soldier's condition. The conflicting audio cue may encourage the user to proceed incorrectly.

Abstract

Mixed Reality (MR) systems are increasingly deployed in safety-critical domains such as medical training and emergency response, where users must interpret automated system guidance while performing complex physical and cognitive tasks. In these high-stakes environments, misleading system feedback can disrupt user engagement and degrade task performance. Such disruptions may occur when incorrect guidance is presented through different interface modalities, including textual instructions, visual overlays, or spoken prompts, which can confuse users, increase cognitive load, and interfere with correct procedural execution. However, existing research lacks comprehensive resources that systematically vary these deceptive guidance modalities and their intensity to study their behavioral impact. To address this gap, we introduce an MR dataset and targeted case study that captures user behavior under deceptive system feedback during a high-stress, safety-critical Tactical Combat Casualty Care training scenario. We collected behavioral telemetry from 29 participants exposed to misleading textual, visual, and auditory interface cues while interacting with the MR system, recording synchronized eye gaze trajectories, field-of-view traces, and detailed behavioral interaction logs. The dataset also includes comprehensive psychometric profiling from 17 validated instruments that measure baseline traits such as self-efficacy and locus of control, together with continuous subjective workload and presence scores. Statistical analysis reveals a highly significant behavioral adaptation effect ($p < .001$), indicating that participants progressively learned to filter misleading MR guidance and recover task efficiency over time. To establish a predictive baseline, we evaluated several machine learning models. The results show that an Explainable Boosting Machine classifier achieves more than 80% accuracy and a micro area under the curve score of 0.92 for detecting engagement degradation using continuous gaze and interaction signals. By releasing this open dataset together with strong predictive baselines, this work provides a valuable resource for studying engagement dynamics and developing anomaly-aware models for immersive systems.

Index Terms: Engagement, Mixed Reality, Cognitive Load, Behavioral Modeling, Temporal Adaptation

1 Introduction

Mixed Reality (MR) systems are increasingly used in safety-critical domains, including defense and medical training, industrial maintenance, and emergency response [26]. In these settings, users rely on digital guidance while performing perceptual, cognitive, and physical tasks [5]. MR platforms can deliver text, audio, visual overlays, and interactive prompts directly within the user's view [44].

*nahmed25@students.kennesaw.edu, †mhassa27@students.kennesaw.edu, ‡wu.p@northeastern.edu, §kzh529@psu.edu, ¶mintrrey@psu.edu, ‖joann.archer@designinteractive.net, **john.kwasinski@designinteractive.net, ††peyton.bailey@designinteractive.net, ‡‡jessyca.derby@designinteractive.net, §§tlan@gwu.edu, ¶¶binli@psu.edu, ‖gtan@psu.edu, ***m.imani@northeastern.edu, †††rislam11@kennesaw.edu.

Although this supports contextual training and operational decision making, it also creates risk when users treat misleading or adversarial system output as credible under time pressure [9]. This concern is especially important in safety-critical scenarios, where small errors in system guidance can affect task execution, workload, and performance [10]. Misleading text, contradictory audio, or incorrect visual placement may disrupt emergency care, industrial work, or military procedures by interfering with attention, decision making, and procedural accuracy [32]. Therefore, it is important to examine how misleading MR information affects engagement, cognitive state, task efficiency, and behavioral adaptation over time [35].

Prior research in human automation interaction has shown that users do not respond to unreliable systems in a uniform way [1, 47]. Their behavior depends on how information is presented, how often errors occur, and how quickly they can recognize inconsistencies in system output [30]. Research has also shown that user engagement, workload, presence, and performance are shaped by interface design, multimodal feedback, and environmental coherence [23]. In addition, research on misinformation and adversarial interfaces suggests that misleading content can affect attention, compliance, and cognitive load in ways that depend strongly on modality and context [33]. However, despite this growing body of work, very few datasets systematically manipulate deceptive information in MR while collecting synchronized gaze, field of view, behavioral logs, task performance measures, and validated self-report instruments. Most existing MR datasets focus on normal interaction conditions rather than controlled deceptive or adversarial system feedback [38].

This limitation makes it difficult to answer several important questions. How do users adapt when misleading information persists throughout an MR task? Do different attack modalities produce different behavioral effects? Can engagement degradation or behavioral disruption be detected directly from high-frequency gaze and interaction signals? How does user performance change over time as participants encounter repeated false information? Answering these questions requires a dataset that combines controlled attack conditions with rich multimodal data and subjective assessments collected during a realistic MR scenario [6].

To address this gap, we present *Believe It. Engage It.*[1], a mixed reality dataset collected in a HoloLens 2 Tactical Combat Casualty Care (TCCC) training scenario under controlled deceptive system feedback. Participants completed the same training task while being exposed to either text-only or multimodal false information at low, medium, or high intensity. Across these conditions, we recorded synchronized gaze trajectories, head and field of view traces, interaction logs, step-level task performance, and a broad set of validated psychometric and self-report measures. This design allows analysis of how users behave under persistent misinformation and how these effects vary by modality and intensity.

The main contributions of this work are as follows:

- We introduce an MR dataset collected in a HoloLens 2 safety-critical training scenario with controlled deceptive system feedback. The dataset includes synchronized gaze traces, field of view logs, interaction streams, task performance records, and validated self-report measures.
- We design a controlled experimental framework that varies both the modality and intensity of misinformation through text-based and multimodal false guidance conditions, enabling systematic analysis of how deceptive information affects user engagement and behavior in MR systems.
- We develop machine learning models to predict user engagement from multimodal behavioral and system data and apply explainability analysis to identify the most influential signals that contribute to engagement changes during deceptive MR interactions.

Preliminary analysis reveals several notable patterns. Participants demonstrated strong temporal adaptation during the MR task. Physiological analysis shows a significant increase in oculomotor strain after exposure ($p < .003$) while nausea and disorientation remained low, confirming the ecological validity of the environment. In addition, baseline engagement models achieved over 80% classification accuracy and a micro-AUC of 0.92 using gaze and interaction data.

2 Background & Related Studies

This section reviews prior research related to immersive systems used for safety-critical training, human interaction with MR guidance, and the behavioral effects of misleading system information.

2.1 MR Systems for Safety Critical Training

MR systems are increasingly used for training in safety-critical domains such as construction, industry, healthcare, and emergency response because they allow users to experience hazardous scenarios without real-world risk [19]. Prior work shows that MR simulators improve learning by enabling trainees to practice complex or rare events that cannot be safely reproduced in real settings, leading to improvements in skill development, knowledge retention, and behavioral performance compared with traditional instruction-based training [14, 24]. In healthcare and emergency response contexts, MR environments support procedural rehearsal, situational decision making, and hands on task execution under realistic operational constraints, while similar benefits have been observed in industrial and construction training where immersive systems improve hazard recognition and compliance with safety procedures [11, 52]. These systems combine spatial overlays, interactive guidance, and multimodal feedback to support situational awareness and sustained engagement during task execution [7]. However, the strong reliance on system-generated guidance also introduces potential vulnerabilities if the information becomes misleading or inconsistent, requiring users to recalibrate trust while continuing to perform under pressure [16].

2.2 Adversarial and Misleading Information in Immersive Interfaces

Immersive interfaces can be disrupted by misleading or contradictory information embedded in the environment [2]. Such disruptions may arise from interface design errors, sensing failures, algorithmic misclassification, poor content generation, or deliberate cognitive attacks [42]. In these situations, systems may still present guidance that appears credible even when the underlying information distorts user perception, attention, or decision-making. Prior work on misinformation shows that the behavioral impact of misleading information depends strongly on how it is presented, including modality, timing, realism, and contextual consistency [34]. In immersive environments, misleading cues may appear as textual prompts, visual overlays, auditory signals, spatial inconsistencies, or multimodal combinations of these elements [4]. Because immersion relies on plausibility and coherence across sensory channels, violations of these expectations can increase cognitive load, disrupt task flow, and reduce user engagement [45, 54]. Over repeated exposures, users may adapt by ignoring system guidance, relying more on prior knowledge, or selectively attending to certain cues, highlighting the complex interaction between system reliability and human behavior in immersive settings. Recent work by Gharaibeh et al. also provides an ontology-driven framework for studying immersive deception. Their validation shows that adversarial actions, such as physiological gaslighting and targeted disinformation, can affect user motor

[1]Dataset request webpage: `https://www.hix-lab.com/dataset-request`. Until the webpage is available, requests may be submitted via the HiX Lab Dataset Request Form. Code and models repository: `https://github.com/Nasim-Ahmed-CS/Believe-It-Engage-It`. For questions regarding the dataset, please contact the corresponding author or visit `https://www.hix-lab.com/`.

behavior and compliance, highlighting the need for empirical tools to evaluate secure immersive interactions [15].

2.3 Datasets for Modeling User Behavior in Immersive Environments

Several datasets have been introduced to study user behavior and physiological responses in immersive environments. Prior work has explored specific signals such as head motion, gaze patterns, and multimodal physiological data during VR interaction. For instance, Wu et al. [51] released a head-tracking dataset for users watching spherical VR videos, while BehaVR analyzes motion and sensor data across multiple applications to study user identification risks in VR [22]. Other datasets focus on affective and cognitive states, including VREED, which captures eye tracking and physiological responses during emotion elicitation [48], and VRMN-bD, which records multimodal signals to study fear responses in interactive VR games [53]. The Mazed and Confused dataset further provides multimodal telemetry collected during real walking tasks in VR [41]. In addition, the SET dataset introduced by Islam et al. [21] combines physiological signals, eye tracking, and head motion to study early forecasting of cybersickness using multimodal deep learning models. While these datasets enable modeling of user state, attention, and behavior in immersive systems, they are largely collected under normal interaction conditions. Consequently, limited data exists for understanding how users behave when immersive systems present misleading or unreliable guidance, motivating the need for datasets that capture behavioral telemetry and subjective measures under controlled deceptive system feedback in MR scenarios.

3 Data Collection and Experimental Design

3.1 Task Description

The TCCC testbed used in this work instructs users to apply a tourniquet in a simulated battlefield scenario, requiring independent decision-making in care under fire and tactical field care settings. The system is built upon an established MR training platform[2] designed for high-intensity, scenario-based medical training. It has been used in operational training contexts and evaluated through expert feedback and training effectiveness assessments. For this study, the testbed was deployed on a Microsoft HoloLens 2 device.

3.2 Task Conditions

The testbed incorporated a cognitive attack based on false information, which has been shown to influence user behavior and reliance on system guidance in immersive systems [49]. User responses in MR environments are shaped by perceived reliability, informational coherence, and spatial consistency, and prior work shows that presentation modality can affect how users interpret system guidance [39, 49]. To examine these effects, false information was injected into the MR interface through either text-only or multimodal cues. Participants were assigned to one of six conditions defined by attack modality and intensity, where modality included text-only or multimodal guidance and intensity levels were Low, Medium, or High. The frequency of injected false information was varied according to models of cumulative reliability degradation, which suggest that repeated reliability violations progressively alter how users rely on and respond to system guidance [17, 20]. All participants completed the same MR TCCC training scenario in which they applied a tourniquet to treat a simulated hemorrhage while interacting with instructional overlays, animations, and task activities involving both a virtual and physical manikin. Each session lasted approximately 20 to 30 minutes, during which participants encountered false information according to their assigned modality and intensity condition, as summarized in Table 1.

[2] https://designinteractive.net/augmed-3/

3.2.1 False Information - Text Only

Text-only insertions were designed to degrade the visual interface without altering the spatial environment. These attacks included rendering pure gibberish across instructional panels, incorrect formatting (e.g., randomized indentations), and simulated error messages directly within the text fields (e.g., rendering an Error 404 prompt). Examples of these text-only attacks are shown in Figure 2(a,b), where the instructional interface presents unreliable text and false system error cues. These textual anomalies forced users to proceed without reliable written guidance. Text-based anomalies such as gibberish, formatting irregularities, and simulated system errors were selected to induce reliability violations without altering core medical instruction. Minor interface errors have been shown to degrade perceived system competence and reduce reliance in technology-mediated environments [17, 20]. In MR contexts specifically, semantic or structural inconsistencies undermine informational credibility and user confidence in system outputs [49].

3.2.2 False Information - Multimodal

In the multimodal condition, false information extended beyond text anomalies to include activity, visual, and auditory inconsistencies. Participants encountered nonsensical text prompts unrelated to the medical procedure (e.g., "Insert Cash or Select Payment Type") together with activity disruptions such as repeated task steps, which increase extraneous cognitive load and interrupt task flow [46]. Figure 2(c,d,e) illustrates multimodal attacks through duplicated assets, misleading spatial cues, and contradictory audio guidance. Visual inconsistencies were introduced through three-dimensional asset anomalies, including misaligned objects, duplicated equipment, and incorrect casualty orientation, which disrupt spatial coherence and reduce perceived system reliability in MR environments [39]. In addition, contradictory auditory cues, such as gunfire continuing after the system indicated that the area was safe, were used to create perceptual conflicts. Prior work shows that violations of environmental plausibility and multimodal consistency can reduce immersion, alter user trust, and affect engagement during MR interaction [43, 49]. These multimodal perturbations were therefore designed to simulate realistic reliability violations while allowing systematic analysis of user behavior under deceptive MR guidance.

3.2.3 Trial Structure

Before data collection, participants completed a short familiarization phase with the HoloLens 2 interface. Each participant then completed one continuous trial using a between-subjects design. As shown in Figure 3, the TCCC scenario included 30 task steps. The 2-minute intervals in Figure 3 represent rolling windows used to aggregate telemetry and collect brief subjective responses. These prompts were given during natural breaks between task steps to reduce disruption to the medical task.

3.3 Participant Details

A total of 30 participants were recruited for this study. One participant was removed due to incomplete telemetry data, resulting in a final sample of 29 participants. The participants were assigned to two conditions: Text Only with 15 users and Multimodal with 14 users, with users distributed across Low, Medium, and High attack intensity levels. The participants ranged in age from 19 to 32 years, with 25 males and 4 females in the sample. The group included individuals from diverse ethnic backgrounds, such as Asian, White, Hispanic or Latino, Black or African American, and multiracial. Educational backgrounds ranged from Associate- to master 's-level programs. Prior experience with augmented reality also varied widely, from participants with no previous exposure to those with more than 100 sessions. This diversity helps ensure the dataset captures a broad range of user behaviors and experience levels. It is important to note that participants were not military or medical experts. This

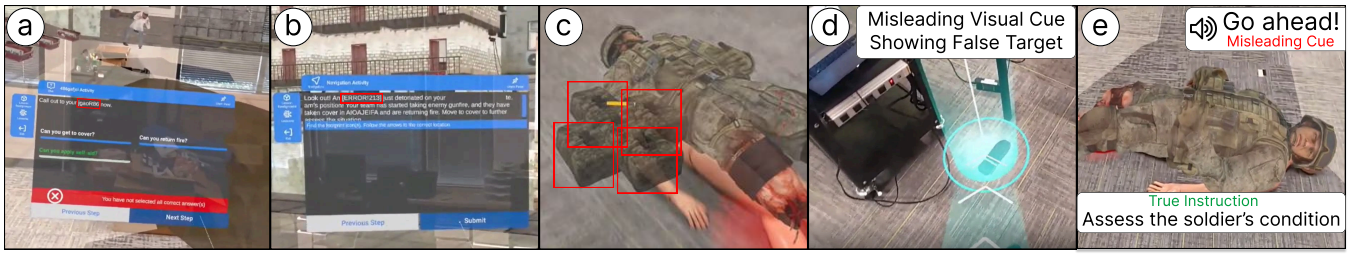

Figure 2: Representative examples of deceptive MR attack conditions. (a,b) Text-only attacks introduce misleading instructional content, including gibberish and false errors. (c,d,e) Multimodal attacks combine visual, activity, and auditory inconsistencies, including duplicated assets, misleading cues, and contradictory audio guidance.

Table 1: Injected false information levels across multimodal and text-only conditions. Frequency indicates how often perturbations occur during the scenario.

Intensity	Multimodal (Text, Activity, 3D, Sound)	Text Only
Low	Text insertions (e.g., gibberish, formatting errors). Frequency: every 3–5 steps.	Text insertions (e.g., gibberish, formatting errors). Frequency: every 3–5 steps.
Medium	Text + activity insertions (e.g., duplicate completed task). Frequency: every 2–4 steps.	Text insertions. Frequency: every 1–3 steps.
High	Text + activity + 3D/sound insertions (e.g., conflicting cues, extra 3D assets). Frequency: every 1–3 steps.	Text insertions. Frequency: every 1–2 steps.

is relevant because they may have relied more on system guidance during the task. Therefore, this sample provides a useful baseline for studying trust and behavioral adaptation in immersive safety-critical scenarios.

3.4 Hypotheses

The primary objective of this data collection is to examine how users physically and cognitively respond to persistent attacks in safety-critical MR environments. Drawing on prior work in human automation interaction, immersive system resilience, and cognitive load theory, we formulate the following hypotheses for the MR testbed.

H1: Physiological Strain.
Participants will experience a measurable increase in oculomotor strain due to the visual demands imposed by the MR interface, while severe nausea or disorientation will remain limited, preserving the ecological validity of the simulation.

H2: Behavioral Adaptation.
Participants will demonstrate temporal adaptation and behavioral resilience to persistent spatial and textual anomalies, reflected in a significant reduction in average task completion time as exposure progresses, indicating that users learn to filter misleading information during task execution.

H3: Engagement Detection.
High frequency gaze and visual signals can serve as reliable indicators of engagement, enabling machine learning models to detect engagement degradation during cognitive attacks in MR environments.

4 Data Description

We collected a comprehensive array of behavioral and physiological data. This provides essential elements for investigating users' engagement, cognitive states, and behavioral resilience.

4.1 Interaction and Behavioral Streams

During each session, the MR testbed continuously recorded participant interaction events and behavioral activity within the training environment. These logs capture how users progressed through the TCCC scenario while responding to system guidance and injected false information. Interaction streams include timestamps for task

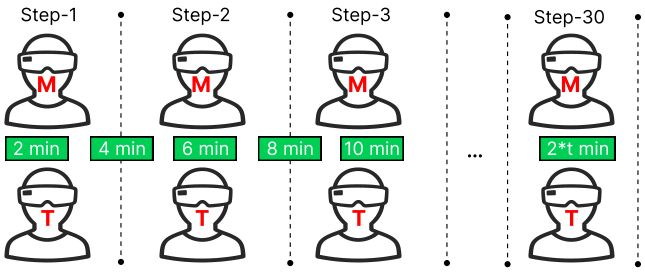

Figure 3: Overview of a session in the TCCC simulation. Participants complete 30 consecutive steps, while subjective data is collected every 2-minute intervals.

steps, activity transitions, and user responses during navigation, selection, and decision-making activities. Each activity also contains metadata describing the task type, step duration, number of attempts, and response correctness. All interaction events are synchronized with the HoloLens recording pipeline, enabling step-by-step reconstruction of user behavior and analysis of task performance, adaptation to misleading guidance, and temporal changes in engagement during cognitive attacks.

4.2 Eye Gaze and Field-of-View Traces

The dataset includes high-frequency eye-gaze telemetry captured by the Microsoft HoloLens 2 eye-tracking system. For each frame, the system records three-dimensional gaze origin and gaze direction vectors along with head position and head orientation, providing a continuous representation of visual attention during MR interaction. The dataset also contains field-of-view object traces that record which virtual objects are visible to the participant at each timestamp, including the number of visible objects, their identity, and their world and screen coordinates. Gaze and field-of-view data are synchronized using shared timestamps, enabling analysis of how visual attention shifts across virtual elements during task execution and cognitive attack conditions.

4.3 Visual Scene Features

The dataset also includes visual scene features extracted from the MR interface to characterize the visual properties of the rendered environment during task execution. These features capture temporal

and structural variations in the visual content that may influence user perception, attention, and cognitive load. Specifically, we compute measures such as brightness flicker, contrast, spatial frequency, edge intensity, and histogram of oriented gradients (HOG) descriptors, which quantify changes in visual intensity, texture, and scene structure as instructional overlays and virtual objects appear within the participant's field of view. All visual features were computed per frame and synchronized with gaze and behavioral telemetry through shared timestamps, enabling analysis of how variations in visual scene properties relate to gaze behavior, task progression, and engagement during cognitive attack conditions [12].

4.4 Task Performance

Participant task performance in the MR scenario was organized into several activity types that represent common interaction patterns during training.

Discovery Activity. In this activity, participants explore the virtual environment and identify relevant areas or items needed for the task. The goal is to observe the surroundings and recognize important clues or objects that may be required for the next steps of the procedure.

Locate and Select Activity. In this activity, participants identify the correct physical or anatomical location and select the appropriate target. This task evaluates whether the participant can correctly determine where to apply an action within the scenario.

Multiple Choice Activity. In this activity, participants respond to a decision prompt by selecting one of several possible options. This type of task measures decision-making and knowledge of the correct response or procedure.

Navigation Activity. In this activity, participants move to a specific area or zone within the scenario. The task evaluates whether the participant can correctly navigate the environment and reach the intended location required for the next step of the task, and encourages embodiment.

Timer Activity. In this activity, participants start a timer that measures how long it takes them to complete a task. This evaluates whether the participant can respond and complete the procedure in an adequate amount of time to save the casualty.

4.5 Self-Report Measures

To capture subjective user experience during the MR training session, we collected several validated questionnaires widely used in human-computer interaction and immersive systems research. These instruments measure workload, presence, usability, trust, and engagement during interaction with the system. All questionnaires were administered during the session and aligned with the behavioral telemetry using synchronized timestamps.

4.5.1 NASA Task Load Index (NASA TLX)

Workload during the MR training task was measured using the NASA Task Load Index (NASA-TLX) [18]. NASA-TLX assesses mental demand, physical demand, temporal demand, performance, effort, and frustration on a scale from 0 to 20, with higher scores indicating greater workload. Participants completed the questionnaire after the assigned MR attack condition, and the scores were used to examine perceived cognitive and physical workload.

4.5.2 Presence Questionnaire (PQ)

The sense of presence in the MR environment was measured using the Presence Questionnaire (PQ) [28]. The PQ evaluates the degree to which users feel immersed in the virtual environment. It captures aspects such as realism of the environment, the ability to act within the environment, and the ability to examine surrounding objects and elements. Higher PQ scores indicate a stronger sense of presence. This measure was used to determine whether deceptive system feedback influenced the user's perceived immersion during the MR training scenario.

4.5.3 System Usability Scale (SUS)

System usability was assessed using the System Usability Scale (SUS) [31], a widely used standardized questionnaire for evaluating the perceived usability of interactive systems. The SUS consists of ten items rated on a Likert scale and produces a final score ranging from 0 to 100. Higher values indicate better perceived usability. In this study, SUS scores were used to evaluate how participants perceived the usability of the MR training system under different attack modalities and intensity levels.

4.5.4 Trust in Automation (TIA)

Trust in the MR system was measured using the Trust in Automation (TIA) scale [27]. The TIA questionnaire evaluates several aspects of user trust, including perceived reliability and competence of the system, predictability and understanding of system behavior, familiarity with the system, perceived developer intention, and the user's general propensity to trust automated systems. These dimensions provide a comprehensive view of how users evaluate automated assistance during task execution. The TIA scores were used to analyze how exposure to false information attacks influenced participants' trust in the MR training system.

4.5.5 User Engagement Scale (UES)

User engagement with the MR training system was measured using the User Engagement Scale (UES) [37]. The UES captures several aspects of engagement, including focused attention, perceived usability, aesthetic appeal, and reward or enjoyment during interaction. Participants rated each item on a scale from 1 to 5, where higher values indicate stronger engagement with the system. The overall engagement score and subscale measures were used to examine how attack modality and intensity influenced user involvement and attentional focus during the MR training session.

4.5.6 Simulator Sickness Questionnaire (SSQ)

Potential simulator sickness symptoms during the MR training session were measured using the Simulator Sickness Questionnaire (SSQ) [25]. The SSQ evaluates symptoms across three primary dimensions, including nausea, oculomotor discomfort, and disorientation. Participants completed the SSQ before and after the MR exposure to capture any changes in physical discomfort associated with the immersive environment. The resulting scores allow analysis of whether extended interaction with the testbed produced measurable physiological strain during the training task.

4.5.7 Visual Fatigue Questionnaire (VFQ)

Visual fatigue was assessed using the Visual Fatigue Questionnaire (VFQ), a 17-item survey designed to measure symptoms of visual strain experienced during MR interaction. Participants rated how noticeable each symptom was at the time of questionnaire completion using a continuous response scale. The VFQ captures indicators such as eye discomfort, visual blur, and difficulty focusing, providing a subjective measure of oculomotor fatigue associated with extended visual engagement in immersive environments [3].

4.5.8 Locus of Control (LoC)

Locus of Control measures whether individuals attribute outcomes to their own actions or to external factors such as luck or fate [13,36]. It was assessed using a 29-item questionnaire, where higher scores indicate a more external locus of control.

4.5.9 Self-Efficacy Scale (SES)

Participants' feelings of self-efficacy, or confidence they have in their abilities to complete tasks, were measured through a 10-item generalized self-report assessment of self-efficacy [40].

Table 2: Descriptive Statistics of Immersive Experience and Workload Metrics. The table reports the Mean and Standard Deviation (SD) for the primary subjective assessments across both attack modalities.

Metric	Text Only ($n = 15$)		Multi Modal ($n = 14$)	
	Mean	SD	Mean	SD
Presence (PQ)	59.33	8.21	62.21	10.09
Usability (SUS)	65.93	11.26	68.29	19.59
Engagement (UES)	3.07	0.26	3.21	0.58
Total Workload (NASA-TLX)	6.46	3.70	5.90	4.54
Trust in Automation (TIA)	7.87	1.46	8.00	2.25

Table 3: Accuracy for Selected Activities under Multi-Modal and Text Only conditions. Values are reported as mean ± standard deviation in percentage.

Activity Type	Multi Modal Accuracy (%)	Text Only Accuracy (%)
Discovery Activity	100.00 ± 0.00	99.17 ± 3.23
Locate & Select Activity	71.43 ± 19.26	67.13 ± 20.61
Multiple Choice Activity	73.21 ± 25.41	74.44 ± 33.70
Navigation Activity	71.43 ± 25.68	85.67 ± 21.78

4.5.10 Embedded Figures Test

Participants' field dependence-independence was measured with a 6-item test. This test assessed how people process information relative to their environment and whether they are able to distinguish individual shapes hidden in a larger, complex figure [50].

4.5.11 Combat Lifesaver (CLS) Knowledge Test

Participants completed a 29-question Combat Lifesaver (CLS) Knowledge Test before using the system [29]. The test assessed baseline knowledge of battlefield medical procedures, including tourniquet use and casualty evaluation. Scores represented the total number of correct answers.

5 Data Characterization

5.1 Descriptive Statistics

To establish a comprehensive behavioral profile of the users under cognitive attack, we first analyzed the subjective experience metrics. Table 2 details the descriptive statistics for Presence, Usability, Engagement, Workload, and Trust across both attack modalities. Notably, the high means for Presence and Usability across both conditions, despite the severe spatial anomalies in the multi-modal attacks, demonstrate the persistence of immersion, indicating that users are dangerously likely to overlook MR system vulnerabilities.

We validated the recording integrity across both experimental conditions. The Text-Only dataset contains a total of **342,434** gaze points ($\mu = 22{,}829$, $\sigma = 8{,}469$), while the Multi-Modal dataset contains **242,254** gaze points ($\mu = 17{,}304$, $\sigma = 5{,}712$). The disparity in data volume (Figure 4) reflects the documented *Missing Data* limitation, primarily attributable to intermittent file write interruptions on the standalone device during specific trials. Nevertheless, with an average of approximately 20,000 gaze samples per valid participant, the sampling density remains sufficient for reliable reconstruction of visual attention patterns.

5.2 Temporal Adaptation and Recovery Patterns

To examine how task efficiency and user performance changed over the course of the 20 to 30-minute TCCC scenario, we split the timestamped telemetry for each participant into two chronological halves and conducted paired t tests on the average step completion times.

As illustrated in Figure 5, the analysis revealed a highly significant temporal adaptation effect across both attack modalities. In the Text Only condition, participants completed steps significantly faster in the second half of the scenario ($M = 22.46$s) compared to the first half ($M = 64.33$s); $t(14) = -6.899$, $p < .001$, $d_z = -1.78$. A similarly strong adaptation was observed in the Multi-Modal condition, with second-half completion times ($M = 43.67$s) significantly faster than the first half ($M = 72.45$s); $t(13) = -5.737$, $p < .001$, $d_z = -1.53$.

This suggests behavioral resilience. However, because there was no clean baseline without deception, the recovery in task efficiency cannot be attributed only to filtering false information. The observed adaptation may reflect several factors, including practice effects, reduced novelty with the interface, and user adjustment to the injected anomalies. This temporal dynamic highlights the richness of the provided dataset. Researchers cannot rely on static post-exposure surveys to understand MR vulnerabilities. Instead, they must utilize the step-by-step performance logs and continuous gaze vectors provided to model real-time human adaptation. Furthermore, the two conditions show that multimodal MR deception creates a direct conflict between physical space, virtual overlays, and audio instructions. This conflict requires users to process competing sensory cues in real time, which may explain the distinct behavioral responses and increased oculomotor strain observed during the task.

To ensure the MR testbed successfully induced a realistic physiological and cognitive load on participants, we analyzed pre- and post-exposure scores for the Simulator Sickness Questionnaire (SSQ) and Visual Fatigue Questionnaire (VFQ). As detailed in Table 4, paired t-tests revealed a highly significant increase in the Overall SSQ Total score across all participants following the 20–30 minute Combat Lifesaver scenario ($t = 3.20$, $p = .003$, $d_z = 0.59$). A deeper subscale analysis indicates that this strain was overwhelmingly driven by the Oculomotor dimension, with oculomotor strain showing highly significant post-exposure increases globally ($t = 3.78$, $p < .001$), as well as independently within both the Text-Only ($p = .017$) and Multi-Modal ($p = .014$) conditions. Conversely, VFQ and the Disorientation dimension of the SSQ did not yield significant differences. This specific profile of oculomotor strain without severe disorientation is a crucial finding, as it confirms the ecological validity of the HoloLens 2 simulation, demonstrating that the testbed successfully immersed users in a visually demanding MR environment without inducing simulation-breaking nausea, thereby ensuring the dataset reliably captures user behavior under genuine physiological load.

5.3 Individual Differences and Psychometric Profiling

A unique contribution of this dataset is the comprehensive psychological and cognitive profiling of the participants, allowing researchers to model not just how an attack functions, but who is most vulnerable to it. Table 5 details the descriptive statistics for baseline traits, including the Locus of Control, Self-Efficacy Scale (SES), Embedded Figures Test (a measure of field dependence), and the CLS Knowledge Test.

The statistics reveal that both the Text-Only and Multi-Modal groups possessed comparable baseline domain expertise ($M = 16.27$ and $M = 15.14$, respectively) and similarly high levels of self-efficacy. This balanced distribution suggests that any behavioral variances or temporal adaptations observed during the Mixed Reality scenario are primarily driven by the cognitive attacks themselves, rather than underlying disparities in medical knowledge or confidence. By releasing these 17 psychometric instruments alongside the high-frequency HoloLens telemetry, we provide the community with the necessary tools to build highly personalized, trait-aware anomaly detection models.

6 Case Study: Engagement Classification

6.1 Interpretable Engagement Modeling

Engagement Classification. To estimate engagement levels during the MR task, we trained several interpretable machine learning models to classify engagement into three categories: Low, Medium, and High. The models were trained on statistical features derived from synchronized behavioral streams, eye gaze telemetry, and visual scene features extracted from the MR interface. Specifically, we evaluated Explainable Boosting Machines (EBM), Random Forest, Decision Tree, and XGBoost classifiers. These models operate on

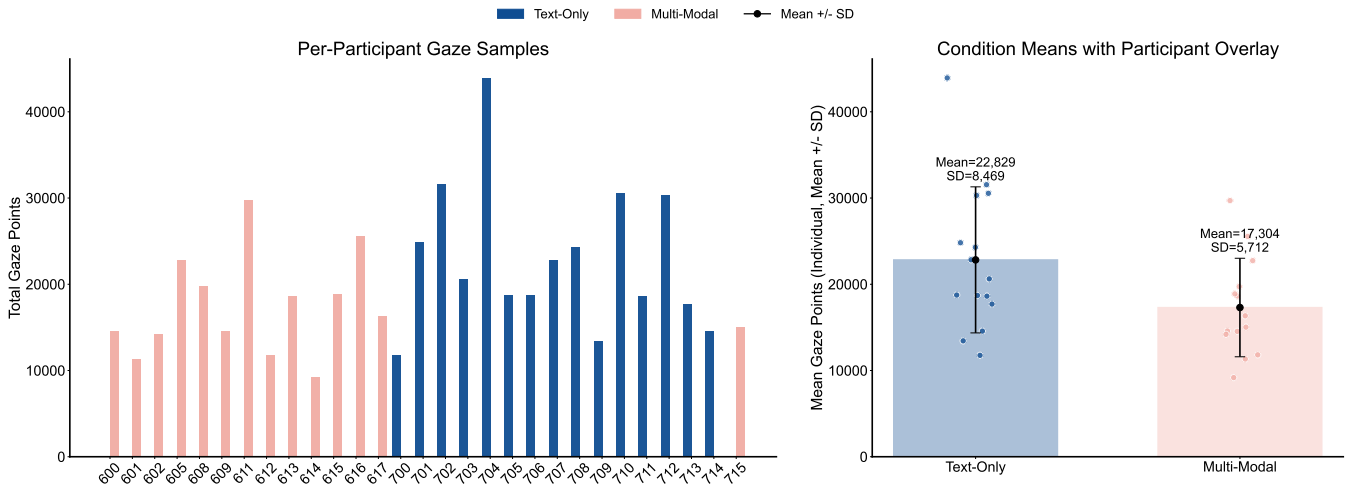

Figure 4: **Data density across experimental conditions.** Left: total gaze samples per participant. Right: condition-level mean gaze samples with participant points overlaid. Black markers and error bars show mean ± SD. Blue indicates Text Only, and salmon indicates Multi-Modal. Text Only shows higher gaze sampling density than Multi-Modal.

Table 4: Pre- and Post-Exposure Physiological and Visual Strain. The table reports paired t-test results for the Simulator Sickness Questionnaire (SSQ) and Visual Fatigue Questionnaire (VFQ). Significant increases in the Oculomotor subscale across all conditions demonstrate the ecological validity and physical demand of the MR testbed.

Measure	Condition	Pre Mean	Post Mean	t-value	p-value	Cohen's d_z
SSQ Total	Overall ($n = 29$)	1.03	4.90	3.20	**.003****	0.59
	Text-Only ($n = 15$)	1.00	6.48	2.71	**.017***	0.70
	Multi-Modal ($n = 14$)	1.07	3.21	1.85	.088	0.49
SSQ Oculomotor	Overall	0.52	6.01	3.78	**.001*****	0.70
	Text-Only	1.01	7.58	2.69	**.017***	0.70
	Multi-Modal	0.00	4.33	2.83	**.014***	0.76
SSQ Nausea	Overall	1.32	3.29	1.80	.083	0.33
	Text-Only	0.64	4.45	2.45	**.028***	0.63
	Multi-Modal	2.04	2.04	0.00	1.000	0.00
SSQ Disorientation	Overall	0.96	2.40	1.36	.184	0.25
	Text-Only	0.93	3.71	1.87	.082	0.48
	Multi-Modal	0.99	0.99	0.00	1.000	0.00
VFQ Visual Fatigue	Overall	18.66	19.10	1.04	0.308	0.19
	Text-Only	17.80	18.53	1.03	.322	0.27
	Multi-Modal	19.57	19.71	0.30	.770	0.08

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

Table 5: Baseline psychometric traits. Descriptive statistics are reported as Mean ± SD for each attack modality.

Psychometric Measure	Text Only (n = 15)	Multi Modal (n = 14)
Locus of Control	10.20 ± 3.63	11.50 ± 4.16
Self-Efficacy Scale (SES)	34.27 ± 4.10	33.29 ± 4.38
Embedded Figures Test (Cognitive Style)	15.47 ± 1.92	14.21 ± 3.29
CLS Knowledge Test (Domain Expertise)	16.27 ± 4.80	15.14 ± 3.13

Table 6: Engagement Regression Model Comparison (Mean ± SD across folds).

Model	MAE ↓	RMSE ↓	R^2 ↑	Pearson r ↑
XGBoost	**0.5240 ± 0.0186**	**0.7293 ± 0.0354**	**0.4833 ± 0.0425**	**0.7027 ± 0.0337**
Random Forest	0.5301 ± 0.0244	0.7353 ± 0.0375	0.4745 ± 0.0485	0.6948 ± 0.0374
Decision Tree	0.5480 ± 0.0288	0.7670 ± 0.0535	0.4279 ± 0.0688	0.6693 ± 0.0477

aggregated descriptors computed from two-minute temporal segments of the interaction timeline, allowing the models to capture behavioral dynamics, gaze patterns, and variations in visual scene properties associated with engagement changes.

Continuous Engagement Estimation: In addition to classification, engagement was modeled as a continuous variable using regression models including Random Forest, Decision Tree, and XGBoost regressors. These models predict engagement intensity directly from the multimodal behavioral and visual signals recorded during the MR interaction. Because of its additive and interpretable structure,

Table 7: Three-class engagement classification performance (Mean ± SD across folds).

Model	Accuracy	Balanced Accuracy	Macro-F1	Micro-AUC
XGBoost	0.7988 ± 0.0309	0.7989 ± 0.0315	0.7959 ± 0.0305	**0.9365 ± 0.0287**
Random Forest	0.7955 ± 0.0440	0.7962 ± 0.0439	0.7925 ± 0.0446	0.9190 ± 0.0372
Decision Tree	0.6999 ± 0.0569	0.7003 ± 0.0576	0.6990 ± 0.0543	0.7750 ± 0.0426
EBM	**0.8056 ± 0.0647**	**0.8059 ± 0.0648**	**0.8041 ± 0.0655**	0.9270 ± 0.0309

Table 8: Three-class engagement classification performance using time-independent models. These models do not use task duration.

Model	Accuracy	Balanced Accuracy	Macro-F1	Micro-AUC
XGBoost	0.7541 ± 0.0382	0.7538 ± 0.0385	0.7512 ± 0.0379	0.8914 ± 0.0321
Random Forest	0.7126 ± 0.0512	0.7119 ± 0.0518	0.7093 ± 0.0524	0.8355 ± 0.0411
Decision Tree	0.6082 ± 0.0714	0.6085 ± 0.0722	0.6071 ± 0.0705	0.7015 ± 0.0558
EBM	0.7377 ± 0.0681	0.7373 ± 0.0685	0.7416 ± 0.0692	0.8722 ± 0.0345

the EBM model was used only for the classification setting, where it provides transparent insight into how gaze behavior, interaction patterns, and visual scene features contribute to engagement prediction [8].

Quartile-based label construction. For each participant timeline, engagement scores were first smoothed using a centered rolling mean with window size $w = 3$:

$$\tilde{s}_i = \text{RollingMean}_{w=3}(s_i). \quad (1)$$

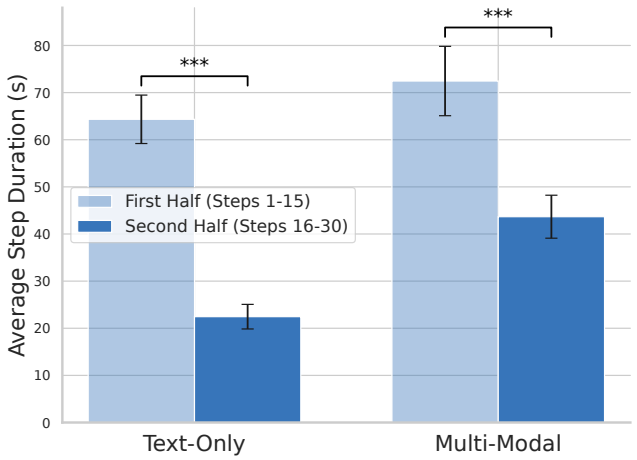

Figure 5: Average step duration (seconds) across the first and second halves of the TCCC scenario for Text Only ($n = 15$) and Multi-Modal ($n = 14$). Error bars show standard error. A significant temporal adaptation effect ($^{***}p < 0.001$) indicates recovery of task efficiency.

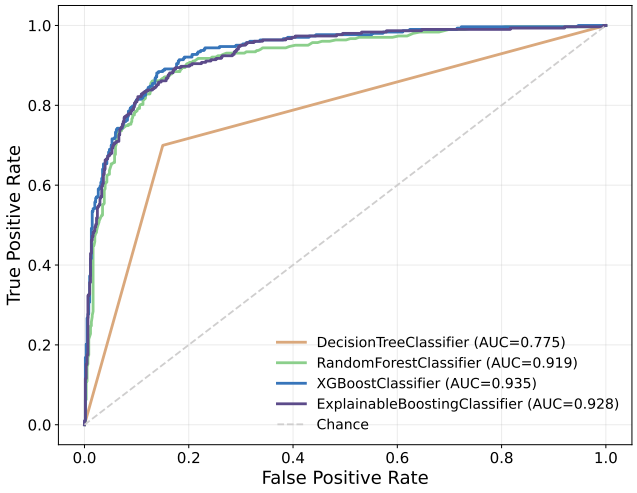

Figure 6: Micro averaged ROC curves for engagement classification models.

The smoothed scores were then ranked and partitioned into quartiles ($q = 4$). Engagement labels were constructed as:

$$Q1 \to \text{Low}(0), \quad Q2, Q3 \to \text{Medium}(1), \quad Q4 \to \text{High}(2).$$

Equivalently, the mapping can be written as

$$y_i = \begin{cases} 0, & q_i = 0 \\ 2, & q_i = 3 \\ 1, & q_i \in \{1,2\}. \end{cases}$$

Feature representation. For each temporal sequence feature, we computed a compact interpretable representation:

$$\{\mu, \sigma, \min, \max, \text{trend}, \overline{|\Delta|}, \max|\Delta|\}$$

representing mean, standard deviation, minimum, maximum, temporal trend, average absolute change, and maximum absolute change across the window. Additional contextual features included condition one-hot indicators and the interaction minute index.

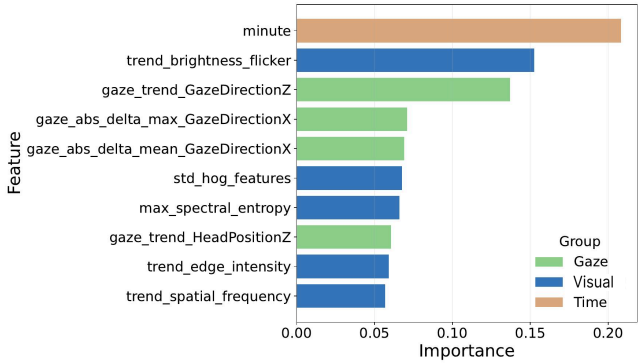

Figure 7: Top predictive features for engagement classification using EBM model.

6.2 Training and Evaluation Protocol

All models were evaluated using 5-fold cross-validation. The folds were participant-independent, so samples from the same participant were not included in both training and testing. In each iteration, four folds were used for training, and one fold was used for testing. This process was repeated five times, with each fold used once as the test set. The final performance was computed as the mean and standard deviation across the five folds. For regression models, we reported MAE, RMSE, R^2, and Pearson correlation. For classification models, we reported Accuracy, Balanced Accuracy, Macro F1 score, and Micro AUC.

6.3 Predictive Performance

The predictive performance of the engagement regression models is summarized in Table 6. Three models were evaluated, including XGBoost, Random Forest, and Decision Tree. Results are reported as the mean and standard deviation across five cross-validation folds. XGBoost achieved the lowest prediction error with MAE of 0.5240 ± 0.0186 and RMSE of 0.7293 ± 0.0354. The model also produced the highest correlation with the ground truth engagement values with $R^2 = 0.4833 \pm 0.0425$ and Pearson correlation $r = 0.7027 \pm 0.0337$. Random Forest produced similar performance with MAE of 0.5301 ± 0.0244, RMSE of 0.7353 ± 0.0375, $R^2 = 0.4745 \pm 0.0485$, and Pearson correlation $r = 0.6948 \pm 0.0374$. The Decision Tree model produced higher error with MAE of 0.5480 ± 0.0288 and RMSE of 0.7670 ± 0.0535. The performance of the engagement classification models is summarized in Table 7 and illustrated in Figure 6. Four models were evaluated, including XGBoost, Random Forest, Decision Tree, and Explainable Boosting Machine. To examine whether engagement classification depends only on task duration, we also evaluated time-independent models after removing the minute index variable. As depicted in Table 8, the EBM model still achieved an Accuracy of 0.7377, Balanced Accuracy of 0.7373, Macro F1 of 0.7416, and Micro AUC of 0.8722. Since random chance is 0.3333 in our three-class setting, this result suggests that spatial gaze and head position features provide meaningful predictive information beyond time. For comparison, the time-inclusive XGBoost achieved an accuracy of 0.7988 ± 0.0309, Random Forest achieved 0.7955 ± 0.0440, and Decision Tree achieved 0.6999 ± 0.0569. Figure 6 reports time inclusive micro averaged AUC values of 0.9365 ± 0.0287 for XGBoost, 0.9190 ± 0.0372 for Random Forest, and 0.7750 ± 0.0426 for Decision Tree.

6.4 Feature Attribution Analysis

The most influential features used by the engagement models include temporal, gaze, and visual scene features statistics. Figure 7 shows the top predictive features for engagement classification using EBM. The interaction minute index appears as the most important feature with an importance value of approximately 0.21. However,

the model also relies on several time-independent gaze and visual features, suggesting that engagement prediction is not driven only by elapsed time. Several gaze-related variables appear among the highest-ranked predictors, including the temporal trend of gaze direction in the Z axis, the maximum absolute change in gaze direction in the X axis, the mean absolute change in gaze direction in the X axis, and the temporal trend of head position in the Z axis. Additional system-level visual features include brightness flicker trend, standard deviation of HOG features, maximum spectral entropy, edge intensity trend, and spatial frequency trend.

7 Discussion

This work introduces a comprehensive multimodal dataset designed to evaluate behavioral resilience in mixed reality environments under deceptive system feedback. The dataset includes synchronized behavioral logs, gaze trajectories, field of view traces, and subjective assessments collected from 29 participants interacting with a HoloLens 2 training system. Across all sessions, the dataset contains more than 580,000 gaze samples and detailed interaction records spanning approximately 20 to 30 minutes of immersive activity per participant. The results show that users are not passive victims of cognitive attacks but instead demonstrate clear temporal adaptation during the scenario. Average step completion times decreased substantially during the second half of the task. These findings also show an important difference between misleading information in MR and misleading information on a tablet or voice chatbot. In MR, deceptive feedback can affect the user's main perceptual environment by creating conflicts between the physical world, virtual overlays, and task instructions. This spatial conflict may help explain the oculomotor strain and adaptation patterns observed in our results. In the Text Only condition, the mean step duration decreased from 64.33 seconds to 22.46 seconds, while in the Multi-Modal condition, the mean step duration decreased from 72.45 seconds to 43.67 seconds. Physiological measurements further confirm the ecological validity of the environment. The Simulator Sickness Questionnaire scores show a significant increase in oculomotor strain after exposure with $p = 0.003$, while nausea and disorientation levels remain relatively low. Baseline machine learning models also demonstrate that engagement degradation can be detected reliably from behavioral telemetry. The engagement classification models showed reasonable predictive performance without relying only on task duration, with the time-independent EBM reaching 0.7377 accuracy and a micro AUC of 0.8722, while the time-inclusive XGBoost achieved a micro AUC of 0.9365. Regression models achieved Pearson correlations above 0.70, indicating meaningful predictive relationships between the extracted features and engagement scores. Feature attribution analysis further shows that gaze-related variables and temporal interaction features are among the most influential predictors. Together, these results provide a strong foundation for future research on resilient mixed reality interfaces and anomaly-aware systems for safety-critical environments.

8 Limitations and Future Works

This study has several important limitations that also open opportunities for future research. First, intermittent file write interruptions on the standalone device resulted in partial telemetry loss for a small number of trials, which reduced the total number of usable gaze samples for some participants. Although the dataset still contains more than half a million gaze samples and dense interaction logs, future work should investigate robust learning methods that can handle missing segments of multimodal telemetry. Approaches such as contrastive learning, representation learning, or temporal imputation models could improve resilience when high-frequency sensor streams contain gaps. Second, the validated dataset contains behavioral recordings from 29 participants. Distributing these participants across modality and attack intensity conditions results in small cell sizes, which limits condition-level comparisons. Although the dataset includes rich temporal data, including over 580,000 gaze points, the participant pool remains limited and may not capture the full diversity of user behavior across larger populations. Future studies should expand the dataset with additional participants and longer observation periods to support population-level modeling and stronger statistical generalization. Third, the present evaluation focuses on a single safety-critical medical training scenario involving tourniquet application within the TCCC environment. Although this task provides a realistic and operationally relevant setting, future work should investigate whether the observed behavioral adaptation and engagement patterns generalize to other mixed reality tasks. Expanding the experimental design to include domains such as industrial maintenance, disaster response training, aviation procedures, or collaborative field operations would help determine whether similar behavioral resilience emerges across different cognitive demands and task structures. Another promising direction involves the integration of the psychometric profiling included in the dataset. The repository provides detailed measurements of baseline traits such as self-efficacy, locus of control, field dependence, and domain knowledge. These signals create an opportunity for the research community to develop personalized engagement and anomaly detection models that adapt to individual cognitive styles and trust tendencies. Future work may combine these traits with high-frequency behavioral telemetry to build adaptive mixed reality systems that dynamically respond to user vulnerability, cognitive load, or susceptibility to misinformation. Such models could support the design of resilient, user-aware interfaces for safety-critical MR applications. Finally, given the scale and multimodal detail of this dataset, future work should expand this work into a more comprehensive journal article. This would allow a deeper analysis of behavioral patterns and modeling opportunities without strict page limits.

9 Conclusion

This paper presented a novel multimodal dataset for studying behavioral resilience in mixed-reality environments under cognitive attack. The dataset integrates synchronized behavioral logs, gaze trajectories, field of view traces, and validated self-report measures collected from participants interacting with a HoloLens 2 training system. Our findings show that users actively adapt to spatial and textual anomalies during task execution rather than passively receiving misleading information. Task completion times decreased significantly during the second half of the scenario, indicating that participants learned to filter out false information and maintain task efficiency. Physiological measurements further confirm the ecological validity of the simulated medical training environment. The system produced measurable oculomotor strain that reflects realistic visual demand while avoiding severe nausea or disorientation that could compromise immersion. In addition, the baseline machine learning models demonstrate that engagement degradation can be detected reliably from behavioral telemetry. The models achieved strong predictive performance and highlighted that high-frequency gaze vectors and temporal interaction features serve as informative biomarkers of cognitive engagement during persistent attacks. By releasing synchronized telemetry data with detailed psychometric profiles, this work supports future research on resilient MR interfaces, anomaly detection, and real-time adaptive systems for safer use in medical training, emergency response, and decision support.

Acknowledgement

This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) under grant HR00112420366. During manuscript preparation, Large Language Models and Grammarly were used only for language refinement. All research contributions were developed by the authors, and all figures are original and manually created without AI image generation.

REFERENCES

[1] N. Ahmed, P. Wu, K. Huang, S. Jung, H. Rheem, G. Tan, M. Imani, and R. Islam. Human task performance and associated internal states in extended reality: a systematic review of cognitive, psychophysiological, and physiological dimensions. *Frontiers in Virtual Reality*, 6:1589256, 2025.

[2] N.-M. Aliman and L. Kester. Malicious design in aivr, falsehood and cybersecurity-oriented immersive defenses. In *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pp. 130–137. IEEE, 2020.

[3] A. W. Bangor. *Display technology and ambient illumination influences on visual fatigue at VDT workstations*. PhD thesis, Virginia Polytechnic Institute and State University, 2000.

[4] A. K. Bhowmik. Virtual and augmented reality: Human sensory-perceptual requirements and trends for immersive spatial computing experiences. *Journal of the Society for Information Display*, 32(8):605–646, 2024.

[5] A. Brunzini, A. Papetti, D. Messi, and M. Germani. A comprehensive method to design and assess mixed reality simulations. *Virtual Reality*, 26(4):1257–1275, 2022.

[6] Y. Chandio, N. Bashir, and F. M. Anwar. Stealthy and practical multimodal attacks on mixed reality tracking. In *2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*, pp. 11–20. IEEE, 2024.

[7] J. Chen, K. P. Seng, J. Smith, and L.-M. Ang. Situation awareness in ai-based technologies and multimodal systems: Architectures, challenges and applications. *IEEE Access*, 12:88779–88818, 2024.

[8] Z. Chen, S. Tan, H. Nori, K. Inkpen, Y. Lou, and R. Caruana. Using explainable boosting machines (ebms) to detect common flaws in data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 534–551. Springer, 2021.

[9] R. Cloete, C. Norval, and J. Singh. A call for auditable virtual, augmented and mixed reality. In *Proceedings of the 26th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–6, 2020.

[10] A. Das, A. Rahman, S. T. Hossain, R. Ahmed, and M. Ara. Mixed reality for human–robot teaming to enhance work health and safety in manufacturing industries: A systematic literature review. *Augmented Human Research*, 10(1):11, 2025.

[11] J. E. Dodoo, H. Al-Samarraie, A. I. Alzahrani, and T. Tang. Xr and workers' safety in high-risk industries: A comprehensive review. *Safety Science*, 185:106804, 2025.

[12] R. L. Franz, S. Junuzovic, and M. Mott. A virtual reality scene taxonomy: Identifying and designing accessible scene-viewing techniques. *ACM Transactions on Computer-Human Interaction*, 31(2):1–44, 2024.

[13] B. M. Galvin, A. E. Randel, B. J. Collins, and R. E. Johnson. Changing the focus of locus (of control): A targeted review of the locus of control literature and agenda for future research. *Journal of Organizational Behavior*, 39(7):820–833, 2018. doi: 10.1002/job.2275

[14] P. S. Gasparello, G. Facenza, F. Vanni, A. Nicoletti, F. Piazza, L. Monica, S. Anastasi, A. Cristaudo, and M. Bergamasco. Use of mixed reality for the training of operators of mobile elevating work platforms with the aim of increasing the level of health and safety at work and reducing training costs. *Frontiers in virtual reality*, 3:1034500, 2022.

[15] T. Gharaibeh, A. Teymourian, A. M. Webb, and I. Baggili. Crafting an immersive game and system deception framework. In *2025 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 295–302. IEEE, 2025.

[16] R. A. Hagen, L. Øverlier, and K. Helkala. Human factors in ai-driven cybersecurity: Cognitive biases and trust issues. *Digital Threats: Research and Practice*, 6(4):1–20, 2025.

[17] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, 53(5):517–527, 2011. doi: 10.1177/0018720811417254

[18] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, vol. 52, pp. 139–183. Elsevier, 1988.

[19] S. Hasanzadeh, N. F. Polys, and J. M. De La Garza. Presence, mixed reality, and risk-taking behavior: A study in safety interventions. *IEEE transactions on visualization and computer graphics*, 26(5):2115–2125, 2020.

[20] K. A. Hoff and M. Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434, 2015. doi: 10.1177/0018720814547570

[21] R. Islam, K. Desai, and J. Quarles. Towards forecasting the onset of cybersickness by fusing physiological, head-tracking and eye-tracking with multimodal deep fusion network. In *Proceedings of the 2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 121–130. IEEE, 2022. doi: 10.1109/ISMAR55827.2022.00027

[22] I. Jarin, Y. Duan, R. Trimananda, H. Cui, S. Elmalaki, and A. Markopoulou. Behavr: User identification based on vr sensor data. *arXiv preprint arXiv:2308.07304*, 2023.

[23] G. Karthikeyan and P. Aruna. User experience optimization in immersive technologies: Techniques, challenges, and opportunities. In *2025 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 908–917. IEEE, 2025.

[24] J. K. Kellie. *Enhancing Safety in the Construction Workforce: A Comparative Study of Hands-On and Virtual Reality (VR) Training Methods*. PhD thesis, University of Wisconsin–Stout, 2025.

[25] R. S. Kennedy and J. E. Fowlkes. Simulator sickness is polygenic and polysymptomatic: Implications for research. In *Simulation in Aviation Training*, pp. 167–182. Routledge, 2017.

[26] S. Khanal, U. S. Medasetti, M. Mashal, B. Savage, and R. Khadka. Virtual and augmented reality in the disaster management technology: a literature review of the past 11 years. *Frontiers in Virtual Reality*, 3:843195, 2022.

[27] M. Körber. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association*, pp. 13–30. Springer, 2018.

[28] E. Kukshinov, A. Hatzipanayioti, M. Slater, and A. Steed. Widespread yet unreliable: A systematic analysis of the use of presence questionnaires. *Interacting with Computers*, p. iwae064, 2025. doi: 10.1093/iwc/iwae064

[29] A. Landman, D. de Vries, and O. Binsch. Retention of military combat lifesaving skills during six months following classroom-style and individualized-style initial training. *Military Psychology*, 35(6):590–602, 2023.

[30] C. Lee, G. A. Rincon, G. Meyer, T. Höllerer, and D. A. Bowman. The effects of visual realism on search tasks in mixed reality simulation. *IEEE transactions on visualization and computer graphics*, 19(4):547–556, 2013.

[31] J. R. Lewis. The system usability scale: Past, present, and future. *International Journal of Human–Computer Interaction*, 34(7):577–590, 2018. doi: 10.1080/10447318.2018.1455307

[32] C. Li, Y. Yin, H. Ye, P. Zheng, and S. K. Gupta. A mixed-reality-augmented deep reinforcement learning approach for multi-robot safe motion generation in human-robot collaborative manufacturing cells. *IEEE Transactions on Automation Science and Engineering*, 2025.

[33] B. Liu, L. Ding, S. Wang, and L. Meng. Misleading effect and spatial learning in head-mounted mixed reality-based navigation. *Geo-Spatial Information Science*, 27(2):408–422, 2024.

[34] M. J. Metzger and A. J. Flanagin. Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of pragmatics*, 59:210–220, 2013.

[35] R. R. Mohanty, P. Selly, L. Brenner, S. Vyas, C. R. Nelson, J. B. Moats, J. L. Gabbard, and R. K. Mehta. From discovery to design and implementation: A guide on integrating immersive technologies in public safety training. *IEEE Transactions on Learning Technologies*, 18:387–401, 2025.

[36] S. Nowicki and M. P. Duke. Foundations of locus of control. *Perceived control: Theory, research, and practice in the first*, 50:147–170, 2017.

[37] H. L. O'Brien, P. Cairns, and M. Hall. 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*, 112:28–39, 2018.

[38] Z. Qi, H. Jin, X. Xu, Q. Wang, Z. Gan, R. Xiong, S. Zhang, M. Liu, J. Wang, X. Ding, et al. Head model dataset for mixed reality navigation in neurosurgical interventions for intracranial lesions. *Scientific Data*, 11(1):538, 2024.

[39] P. A. Rauschnabel, R. Felix, and C. Hinsch. Augmented reality marketing: How mobile ar-apps can improve brands through inspiration. *Journal of Retailing and Consumer Services*, 49:43–53, 2019.

[40] R. Schwarzer and M. Jerusalem. The general self-efficacy scale (gse). *Anxiety, Stress, and Coping*, 12(1):329–345, 2010.

[41] J. N. Setu, J. M. Le, R. K. Kundu, B. Giesbrecht, T. Höllerer, K. A. Hoque, K. Desai, and J. Quarles. Mazed and confused: A dataset of cybersickness, working memory, mental load, physical load, and attention during a real walking task in vr. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 1048–1057. IEEE, 2024.

[42] A. Shalaby. Classification for the digital and cognitive ai hazards: urgent call to establish automated safe standard for protecting young human minds. *Digital Economy and Sustainable Development*, 2(1):17, 2024.

[43] M. Slater. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3549–3557, 2009. doi: 10.1098/rstb.2009.0138

[44] M. Speicher, B. D. Hall, and M. Nebeling. What is mixed reality? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–15, 2019.

[45] K. M. Stanney, C. Hughes, H. Nye, E. Cross, C. Boger, and S. Deming. Interaction design for augmented, virtual, and extended reality environments. In *Human-Computer Interaction*, pp. vol4–355. CRC Press, 2024.

[46] J. Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285, 1988.

[47] K. A. Szczurek, R. M. Prades, E. Matheson, J. Rodriguez-Nogueira, and M. Di Castro. Multimodal multi-user mixed reality human–robot interface for remote operations in hazardous environments. *IEEE Access*, 11:17305–17333, 2023.

[48] L. Tabbaa, R. Searle, S. M. Bafti, M. M. Hossain, J. Intarasisrisawat, M. Glancy, and C. S. Ang. Vreed: Virtual reality emotion recognition dataset using eye tracking & physiological measures. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 5(4):1–20, 2021.

[49] G. Taub, A. Elmalech, and N. Aharony. Trust and attitude toward information presented using augmented reality and other technological means. *Heliyon*, 10(4), 2024.

[50] H. Witkin, P. Oltman, E. Raskin, and S. Karp. Manual for embedded figures test, and group embedded test, 1971.

[51] C. Wu, Z. Tan, Z. Wang, and S. Yang. A dataset for exploring user behaviors in vr spherical video streaming. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 193–198, 2017.

[52] O. Zechner, D. García Guirao, H. Schrom-Feiertag, G. Regal, J. C. Uhl, L. Gyllencreutz, D. Sjöberg, and M. Tscheligi. Nextgen training for medical first responders: advancing mass-casualty incident preparedness through mixed reality technology. *Multimodal Technologies and Interaction*, 7(12):113, 2023.

[53] H. Zhang, X. Li, Y. Sun, X. Fu, C. Qiu, and J. M. Carroll. Vrmn-bd: A multi-modal natural behavior dataset of immersive human fear responses in vr stand-up interactive games. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 320–330. IEEE, 2024.

[54] J. Zhou, S. Z. Arshad, S. Luo, and F. Chen. Effects of uncertainty and cognitive load on user trust in predictive decision making. In *IFIP conference on human-computer interaction*, pp. 23–39. Springer, 2017.