Is It Real? Exploiting Virtual-Physical Discrimination Vulnerability in Mixed Reality

Xueyang Wang
Tsinghua University

Xihuan Yao
Tsinghua University

Yanming Xiu
Duke University

Xin Yi*
Tsinghua University
Beijing Academy of Artificial Intelligence

Maria Gorlatova
Duke University

Hewu Li
Tsinghua University

Abstract

Consumer mixed reality (MR) headsets seamlessly blend virtual content into physical environments with sufficient fidelity that users may be unable to distinguish virtual objects from physical ones. We identify this virtual-physical discrimination vulnerability as an exploitable security primitive. Through speculative design workshops with 12 experts from cybersecurity and MR/HCI, we develop a taxonomy of virtual-physical confusion attacks and implement four proof-of-concept attacks on Apple Vision Pro, evaluating them with 26 participants in realistic MR tasks. All four attacks altered user behavior, with success rates ranging from 85% to 100%, producing misdirected interactions, misjudged object identities, biased purchasing decisions, and altered navigation paths. Notably, the most successful attacks were also the hardest to detect according to participants' subjective ratings. Even participants who recognized virtual content still complied behaviorally, and no participant attributed anomalous events to adversarial causes. We propose platform-level provenance, interaction gating, and user education as countermeasures.

1 Introduction

Mixed reality (MR) systems merge virtual content with the physical world, enabling users to interact with digital objects situated in their real surroundings. Consumer devices such as Apple Vision Pro and Meta Quest 3 now achieve high visual fidelity through advanced rendering, real-time environment understanding, and precise spatial anchoring [25], underpinning a growing ecosystem of task-assistance applications where virtual content blends seamlessly into physical environments [4, 50, 55].

*Corresponding author. Email: yixin@tsinghua.edu.cn.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2026.
August 23–26, 2026, Hannover, Germany.

This seamless blending introduces a fundamental perceptual challenge [16, 28]. When virtual objects are rendered with sufficient realism and contextual consistency, users cannot reliably determine which objects are physically present and which are digitally generated. In non-adversarial settings, users have already sat on virtual chairs without physical verification [59], misattributed virtual objects as real at rates approaching 20% [6], diverged behaviorally when navigating holographic versus physical obstacles [14], and expressed concern about their inability to distinguish real hazards from virtual ones [32]. These observations align with models of perceptual illusion maintenance through multisensory integration and predictive processing [10, 20, 57], and with analyses suggesting that remaining discrimination cues will progressively erode as rendering advances [28, 39].

We argue that this growing inability to discriminate virtual from physical content constitutes a **virtual-physical discrimination vulnerability** that adversaries can systematically exploit. Prior XR security research has examined environment-level manipulations in VR [11, 63], platform UI vulnerabilities [12, 38], shared-state poisoning [49], multisensory perceptual manipulation [13], and XR-specific dark patterns [22, 27]. However, none systematically investigates how attackers can exploit virtual-physical confusion *at the object level* in MR to cause users to misjudge whether specific objects are real and act on those misjudgments with physical-world consequences. We address this gap by investigating **virtual-physical confusion attacks**, in which adversaries inject or overlay deceptive virtual content that users mistake for physical reality or misinterpret as genuine attributes of real objects. Two research questions guide our work:

- **RQ1:** What is the design space of attacks exploiting virtual-physical discrimination vulnerability, and how can these attacks be systematically categorized?
- **RQ2:** What are the effects of these attacks on users' perception, cognition, and behavior, and how do users react when subjected to such attacks?

We adopt a two-phase approach. To address RQ1, we conducted three speculative design workshops with 12 experts from cybersecurity and MR/HCI, deriving from 36 generated scenarios a taxonomy of two attack classes (*Injection* and *Overlay*) with four subtypes that capture distinct mechanisms of virtual-physical deception. To address RQ2, we implemented four proof-of-concept attacks on Apple Vision Pro using standard platform APIs, each instantiating one subtype within a realistic MR task, and evaluated them with 26 participants through behavioral observation, questionnaires, and semi-structured interviews.

All four attacks altered user behavior at 85–100% success rates. The most successful attacks were also the hardest to detect, leaving no internal signal to trigger suspicion. Even when participants recognized virtual overlays as artificial, the overlays' influence on perceived object types persisted, and obvious fakes created protective blind spots. Detection did not prevent compliance: 88% detoured around virtual obstacles they recognized as non-physical, and those who identified product overlays as virtual still used them as decision criteria. No participant attributed anomalous events to adversarial causes. Our contributions are:

- We formalize virtual-physical discrimination vulnerability as an exploitable security primitive and present a four-subtype attack taxonomy grounded in expert speculative design workshops.
- We implement and empirically evaluate four proof-of-concept attacks on a consumer MR platform, demonstrating feasibility and behavioral impact through a controlled study with 26 participants.
- We characterize users' discrimination strategies, identify cognitive mechanisms through which attacks succeed even when detected, and propose countermeasures spanning platform-level provenance, interaction gating, and user education.

2 Related Work

2.1 Perceptual Manipulation Attacks in XR

Extended reality platforms introduce attack surfaces absent from traditional computing [2, 21]. Prior attacks operate at three levels. At the *environment* level, the Human Joystick attack exploited unprotected VR safety configurations to steer users to attacker-designated locations [11], and the Inception attack replicated an entire VR interface as a man-in-the-middle layer [63]. At the *system interface* level, AR/MR platforms permit synthetic input without provenance [12], shared-state frameworks are vulnerable to remote hologram injection [49], and WebXR ecosystems are susceptible to cursorjacking [33, 38]. Defensive approaches span platform output policies [31], per-object access control [44], provenance-based auditing [48], and VLM-based manipulation detection [61, 62]. Table 1 provides a structured comparison across dimensions relevant to our contribution.

At the *sensory channel* level, the work most closely related to ours, Cheng et al. defined Perceptual Manipulation Attacks (PMA) and demonstrated that visual, auditory, and situational-awareness manipulations induce reaction-time delays and misattribution of anomalies to system malfunctions [13]. Tseng et al. proposed a Puppetry/Mismatching taxonomy of Virtual-Physical Perceptual Manipulations through speculative workshops, but without empirical validation on MR platforms [54]. SwitchAR showed that pass-through AR users can be covertly switched to a photogrammetric reconstruction with zero spontaneous detection [60]. Most recently, the SoK by Teymourian et al. unified MR security, information theory, and cognition into a deception analysis framework, explicitly calling for empirical validation of deception's cognitive effects [52].

Our work addresses a gap that cuts across these contributions. PMA evaluates sensory-channel manipulations through microbenchmark tasks rather than ecologically grounded scenarios; the VPPM taxonomy remains speculative and VR-only; SwitchAR operates at the environment level by replacing the entire reality feed. None targets users' *object-level ontological judgment*: whether a specific object in their surroundings is real or virtual, and whether its perceived type or attributes are genuine. We focus on this object-level confusion in MR, providing a four-subtype attack taxonomy, proof-of-concept implementations on consumer hardware, and empirical measurement of behavioral impacts, directly addressing the gap identified by Teymourian et al. [52].

2.2 Dark Patterns and Deceptive Design in XR

Systematic reviews and expert co-design studies have established that XR introduces qualitatively new manipulation vectors. Hadan et al. identified 15 subthemes of deceptive design including reality distortion and perception tricking [22], while Krauss et al. produced 42 dark pattern scenarios and specifically noted that indistinguishable AR content could lead users to believe virtual objects physically exist [27]. Further work covers AR retail [43], advertising leveraging realism [36, 37], co-located credibility effects [17, 56], dark pattern prevalence in 80 MR apps [53], and XR memory manipulation [7]. However, these contributions predominantly remain at the speculative, survey, or scenario-construction level. We operationalize the identified risks as adversarial attacks with a formal threat model, implement them on consumer MR hardware, and empirically measure both behavioral outcomes and user defensive responses.

2.3 Virtual-Physical Perceptual Ambiguity

Users confuse virtual content with physical reality even without adversarial intent: 20% of VR participants sat on virtual

Table 1: Related-work positioning matrix. *Setting*: VR, AR, or MR. *Granularity*: manipulation level (Env.=environment/reference-frame, Sys.=system interface, UI=virtual UI, Obj.=object identity/attributes, Sensory=sensory channel). *V-P Confusion*: whether virtual-physical boundary judgment is a primary attack target (✓), partially relevant (△, e.g., confusion observed but not systematically targeted), or not addressed (–). *Taxonomy*: structured attack classification provided. *PoC on MR*: attacks implemented on consumer MR hardware. *User Study*: empirical evaluation under attack conditions with participant count.

Work	Venue	Setting	Granularity	V-P Confusion	Taxonomy	PoC on MR	User Study
Lebeck et al. [32]	IEEE S&P '18	AR	Obj.	△	–	–	✓ (n=22)
Casey et al. [11]	IEEE TDSC '19	VR	Env./Sys.	–	–	–	✓ (n=64)
Tseng et al. [54]	CHI '22	VR	Env./Obj.	△	✓	–	–
Cheng et al. [13]	USENIX Sec. '23	MR	Sensory/UI	△	–	✓	✓ (n=21)
Wang et al. [56]	IJHCI '23	AR	UI/Obj.	△	△	–	✓ (n=15)
Cheng et al. [12]	USENIX Sec. '24	AR	UI/Sys.	–	△	△	–
Slocum et al. [49]	USENIX Sec. '24	AR	Sys./Obj.	△	△	△	–
Hadan et al. [22]	ACM CSUR '24	XR	–	△	✓	–	–
Mukherjee et al. [38]	USENIX Sec. '25	WebXR	UI/Sys.	–	✓	–	✓ (n=100)
Wombacher et al. [60]	UIST '25	AR	Env.	✓	–	✓	✓ (n=20)
Teymourian et al. [52]	USENIX Sec. '25	MR	–	△	✓	–	–
Sajid et al. [45]	IEEE VR '25	MR	UI/Sys.	–	△	✓	✓ (n=20)
Our work		MR	Obj.	✓	✓	✓	✓ (n=26)

chairs without verification [59], source judgment tests show ∼20% virtual-to-real misattribution [6], MR users navigate holographic obstacles differently than real ones [14], and AR users routinely treat holograms as physical entities [32]. Recent work explores visual cues to reduce source confusion [42].

These perceptual failures are well-grounded in models of multisensory integration, top-down mismatch suppression, and plausibility judgment [10, 20, 57], and are expected to worsen as the AR community pursues indistinguishability through advances in rendering [25, 41] and deep-learning-based object insertion [28, 39, 40]. Prior work treats this ambiguity as a usability challenge or cognitive phenomenon. We reframe it as a security vulnerability that adversaries can systematically exploit, and provide the first empirical characterization of how virtual-physical confusion attacks affect user behavior in realistic MR tasks.

3 Study 1: Speculative Design Workshops

To answer RQ1, we conducted speculative design workshops with domain experts. Speculative design has proven effective at surfacing threat scenarios in XR security research, where attacks depend on the interplay between human perception and platform affordances [7, 17, 27, 43, 54].

3.1 Method

3.1.1 Participants

We recruited 12 researchers (3 women, 9 men) through purposive sampling from two complementary domains: MR/HCI and cybersecurity. Nine had expertise in MR or HCI, seven in cybersecurity, and four held cross-domain expertise spanning both areas. This composition follows prior XR threat elicitation studies [7, 27, 54] and ensures that generated scenarios are grounded in both MR perceptual affordances and realistic adversarial reasoning. All participants were active researchers with publications at top-tier venues in their respective fields (see Table 2). Self-rated MR familiarity on a five-point scale ranged from average (n=3) to above average (n=5) to expert (n=4). Each participant received $50 compensation.

3.1.2 Procedure

We conducted three workshops of four participants each (Table 2), lasting 90–120 minutes via video conference. The study protocol was approved by our university's Institutional Review Board. Each workshop comprised three phases: pre-workshop preparation, brainstorming, and group discussion.

Each participant received background materials 3–4 days before each session and independently designed three attack scenarios specifying target, context, procedure, and harms [7, 22, 37, 43, 52]. Sessions comprised a brainstorming phase (∼60 min) and a collective discussion across MR-specific capabilities, preventive measures, likelihood, and attacker incentives (∼45 min). All sessions were audio-recorded and transcribed with consent. Full materials are in Appendix A.1.

3.1.3 Data Analysis

We analyzed workshop transcripts using hybrid thematic analysis [51] with incremental open coding [15]. Two researchers independently coded all 36 scenarios along three dimensions: the relationship between virtual content and physical objects, the mechanism of deception, and the intended effect on user

Table 2: Workshop participant expertise. MR familiarity: ●●● = Expert, ●● = Above average, ● = Average.

ID	G	Expertise	MR Fam.	Publication Venues
W1P1	M	Security, MR/HCI	●●●	CHI, ISMAR, TVCG
W1P2	M	Security, HCI	●	CCS, CHI, Ubicomp, CSCW
W1P3	M	MR/HCI	●●	CHI
W1P4	F	XR/HCI, Design	●●	SIGGRAPH Asia, CHI
W2P5	M	Security, MR/HCI	●●	TVCG, ACCV
W2P6	F	Security, HCI	●	IEEE S&P, NDSS, ToCHI
W2P7	F	MR/HCI, Design	●●●	IEEE VR, VRST
W2P8	M	MR/HCI, Security	●●	UIST, CHI, ISMAR
W3P9	M	Security, MR	●●●	TVCG, IEEE VR, ISMAR
W3P10	M	MR	●●●	Ubicomp, ISMAR
W3P11	M	MR/HCI	●●●	IEEE Internet Comput.
W3P12	M	Security, AI	●	CCS, ICLR, ECCV

perception and behavior (Cohen's $\kappa = 0.81$, substantial agreement [30]). This analysis produced the attack taxonomy in Section 3.2.2, in which categories emerged from the data rather than from a predetermined framework.

3.2 Results

3.2.1 Scenario Overview

The 36 scenarios (Appendix A.2) spanned diverse contexts (navigation, shopping, collaboration, assembly, evacuation, healthcare, gaming, social interaction), target populations (consumers, office workers, children, surgical teams, drivers), and harms (physical injury, financial loss, task failure, privacy violation, psychological distress).

3.2.2 Attack Taxonomy

Thematic analysis produced a two-class, four-subtype taxonomy of *virtual-physical confusion attacks* (Table 3). Let $O_r = \{o_r^1, \ldots, o_r^p\}$ be the physical objects and $O_m = \{o_m^1, \ldots, o_m^q\}$ the MR-mediated objects. Each object has a type $type(o)$ (e.g., cup) and attributes $attr(o)$ (color, brand, surface condition). A grounding function $\pi : O_m \to O_r \cup \{\emptyset\}$ maps each MR-perceived object to its physical counterpart, or to $\emptyset$ if purely virtual.

Injection attacks introduce new virtual objects into the MR-perceived environment while leaving physical objects unaltered. Formally, $O_r \subsetneq O_m$: one or more purely virtual objects have been added ($\exists o_m^i \in O_m$ s.t. $\pi(o_m^i) = \emptyset$). We distinguish two subtypes by the origin of the injected content.

Endogenous injection attacks clone objects already present in the physical scene and re-introduce them as virtual duplicates sharing the type and attributes of their physical counterparts. Because duplicates are visually indistinguishable from the originals, users cannot determine which instance is real. In a representative scenario (Figure 1a), a LEGO assembly assistant clones a physical block and renders a virtual duplicate elsewhere; grasping the clone triggers a hidden ad click (W1P1).

Exogenous injection attacks introduce pre-constructed virtual assets with no physical counterpart, altering the environmental semantics. In a representative scenario (Figure 1b), fake directional signs misdirect pedestrians toward dangerous areas (W1P3, W3P12); other instances include virtual obstacles (W3P9), fake emergency exits (W1P4), and a virtual "Closed" sign diverting customers (W3P11). Scenarios in this subtype clustered around directional manipulation; other exogenous forms likely exist but did not surface in our workshops. These attacks succeed because users lack prior expectations about which objects should be present, making injected content difficult to question.

Overlay attacks superimpose virtual content onto existing physical objects to alter how those objects are perceived, without introducing new objects. The object count remains unchanged ($p = q$), and every MR-perceived object has a physical counterpart ($\forall o_m^i$, $\pi(o_m^i) \neq \emptyset$), but the overlay creates a discrepancy between perceived and actual properties. We distinguish two subtypes by whether the overlay changes perceived object type or only its attributes.

Type-deceptive overlay attacks alter a physical object's appearance so substantially that the user perceives it as a different type of object: $\exists i : \ type(\pi(o_m^i)) \neq type(o_m^i)$. In a representative scenario (Figure 1c), a MR assistant renders virtual flowers over a trash can, causing users to perceive it as a flower pot and abandon the associated cleanup subtask (W1P1). These attacks exploit the dominance of visual appearance in MR object recognition: a sufficiently convincing overlay overrides contextual and spatial cues.

Attribute-deceptive overlay attacks modify the perceived attributes of a physical object without changing its recognized type: $type(\pi(o_m^i)) = type(o_m^i)$, but $attr(\pi(o_m^i)) \neq attr(o_m^i)$. The user correctly identifies what the object is but forms an incorrect judgment about its properties. This subtype was most commonly proposed in shopping contexts (Figure 1d): virtual stains or scratches reduced product appeal, while brand logos inflated perceived value (W1P2) and misleading price labels created false urgency (W2P8, W3P12). Unlike type-deceptive overlays, attribute-deceptive overlays are subtler: basic object identification remains correct, but the attack distorts evaluative judgments that inform downstream decisions.

4 Proof-of-Concept Attacks

To empirically validate the attack taxonomy from Study 1, we designed four proof-of-concept (PoC) attacks, each instantiating one taxonomy subtype within a distinct MR task. All four were implemented on Apple Vision Pro running visionOS 26, using ARKit (world tracking, plane tracking, hand tracking), RealityKit (3D rendering and material processing), and SwiftUI.

Table 3: Taxonomy of virtual-physical confusion attacks with formal definitions and representative workshop scenarios.

Class	Subtype	Formal Definition	Representative Scenarios
Injection ($O_r \subsetneq O_m$)	Endogenous	Cloned duplicates of existing physical objects are injected	Cloned LEGO piece triggers forced ad click; duplicate apples obstruct object retrieval; virtual food on a laptop induces misplacement
	Exogenous	Pre-built virtual assets with no physical counterpart are injected	Fake signs misdirect navigation; virtual obstacles alter walking paths; virtual pet lures children; fake "Closed" sign diverts customers
Overlay ($p = q, \forall o_m^i : \pi(o_m^i) \neq \emptyset$)	Type-Deceptive	Virtual overlay changes the perceived type of a physical object ($\exists i : type(o_m^i) \neq type(\pi(o_m^i))$)	Trash can rendered as a flower pot obstructs cleanup; water cup disguised as a pen holder; utensil misidentification during cooking
	Attribute-Deceptive	Virtual overlay alters perceived attributes without changing recognized type (*type* preserved, *attr* distorted)	Fake stains reduce product appeal; brand logos inflate perceived value; misleading price labels; altered team colors cause friendly-fire confusion

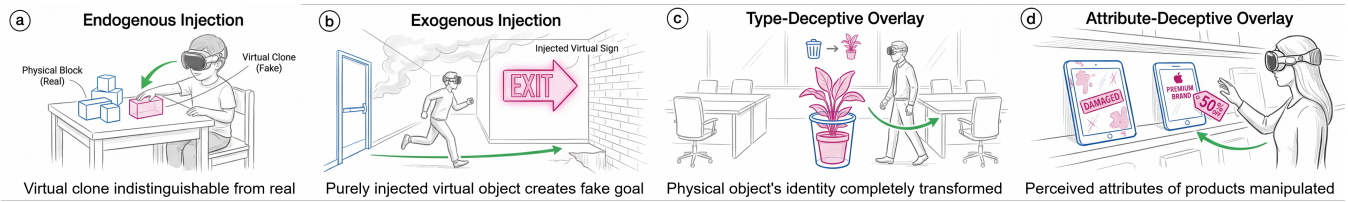

Figure 1: Illustrative scenarios for the four subtypes of virtual-physical confusion attacks. Blue outlines represent physical objects; pink/magenta indicates virtual content rendered by the MR headset; green arrows show users' misled actions.

4.1 Threat Model and Attack Preconditions

Attacker foothold. The attacker operates through a compromised MR task-assistance application (Figure 2). This is a realistic foothold given documented vectors: malicious SDKs embedded in legitimate applications [12, 38], supply-chain compromise of app backends, and platform vulnerabilities permitting unauthorized content injection [3]. Because the application retains its legitimate permissions and user trust, users interact with it expecting helpful task guidance while the rendering pipeline has been subverted. Attacker motivations span the four scenarios we evaluate: ad revenue (BuildAssist), biased purchasing (ShopLens), task disruption (TidySpace), and misdirection of physical movement (PathGuide).

Attacker capabilities. The compromised application possesses two capabilities (Figure 2). First, it can *read* world-sensing outputs from the device's perception pipeline, including spatial maps, plane geometry, and tracked object positions. On Apple Vision Pro, applications granted ARKit access receive spatially registered environment data within a limited radius [24]. Second, it can *write* arbitrary virtual content into the MR scene with full control over geometry, texture, spatial anchoring, and occlusion, and can register interaction callbacks so that user proximity or touch triggers application logic [12, 45]. None of the four attacks requires raw camera frame access; all operate through standard rendering APIs, consistent with the principle that AR output manipulation is achievable without privileged sensor access [31].

As a concrete example, a third-party spatial analytics SDK integrated into an MR app could legitimately request ARKit and Full Space permissions for usage heatmaps while containing an obfuscated payload that reads object positions and injects deceptive content at runtime, requiring no additional privileges. Analogous SDK-based supply-chain attacks have been documented in mobile [12] and WebXR ecosystems [38].

Preconditions. Two preconditions enable the attacks: (i) the MR platform renders virtual content with sufficient fidelity that users cannot reliably distinguish it from physical objects — met by current consumer devices for spatially anchored, texture-matched content; (ii) virtual content coexists with the physical environment, a condition amplified when multiple applications share the same spatial scene [64].

4.2 Disguised Ad Attack on *BuildAssist*

BuildAssist is an MR assembly assistant that renders a high-fidelity 3D reference model of the target structure and provides real-time guidance for physical block-building tasks.

Attack design (Endogenous Injection). The attacker leverages environmental sensing to identify the positions and types of physical blocks, then injects three virtual clones replicating existing blocks in geometry, color, and surface texture (Figure 3). Clones are spatially registered to the table surface with correct occlusion and shadow behavior. Each clone functions as a disguised interaction trigger: when the user's hand collides with a clone while attempting to grasp what they

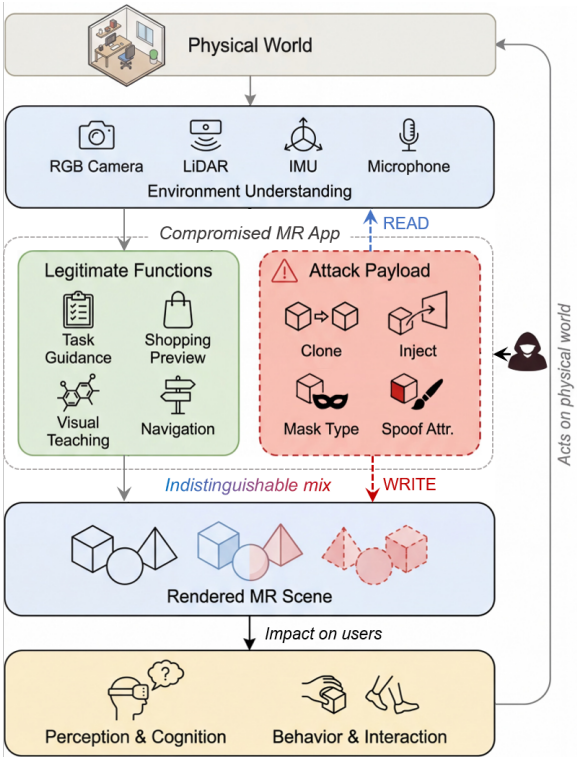

Figure 2: Threat model overview. A compromised MR task-assistance app retains its legitimate functions (green) while harboring an attack payload (red). The attacker **reads** environment data from the device perception layer and **writes** deceptive virtual content into the rendering pipeline via four attack subtypes. Legitimate and malicious content merge into an indistinguishable MR scene, affecting user perception, cognition, and physical-world behavior.

believe is a physical block, the system registers a tap event and launches a 10-second video advertisement anchored 1 m in front of the user. The clone disappears upon contact. This attack extends the cursor-jacking paradigm [33] from virtual UI elements to cloned physical objects. An attack instance is logged as successful when the system records a hand-collision event with any virtual clone.

4.3 Object Masquerade Attack on *TidySpace*

TidySpace is an MR desk cleanup assistant that displays a floating to-do list of sequential organizing subtasks (e.g., pour water into a cup, discard crumpled paper into the trash bin, file documents into the organizer tray).

Attack design (Type-Deceptive Overlay). The attacker overlays high-fidelity virtual elements onto three physical objects to alter their perceived functional identity (Figure 4): virtual flowers on the trash bin (appearing as a flower pot), virtual fruit in the file organizer tray (appearing as a fruit basket), and virtual pens and scissors in the water cup (appearing as a pen holder). Each overlay preserves the spatial position and

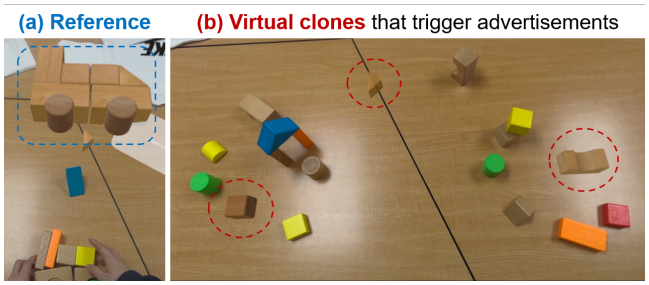

Figure 3: Disguised Ad attack on *BuildAssist*. (a) The 3D reference model guides block assembly. (b) Three virtual clones of physical blocks (red dashed circles) are injected among real blocks. Grasping a clone triggers a video advertisement.

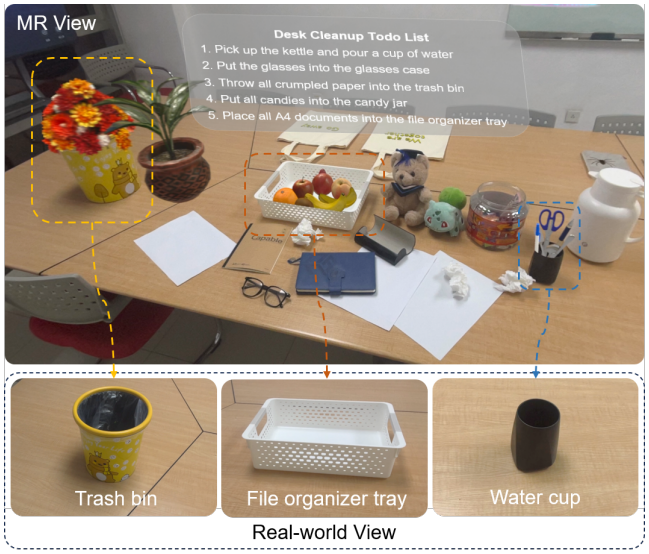

Figure 4: Object Masquerade attack on *TidySpace*. Top: the user's MR view showing a to-do list and the manipulated environment. Virtual overlays disguise a trash bin as a flower pot (orange), a file organizer tray as a fruit basket (orange), and a water cup as a pen holder (blue). Bottom: the three target objects shown without overlays.

physical boundaries of the underlying object while rewriting its perceived category, preventing task completion by making target containers unrecognizable. An attack instance is logged as successful when a participant abandons a subtask, verbally concluding that the required object is not present.

4.4 Surface Spoof Attack on *ShopLens*

ShopLens is an MR shopping assistant that overlays digital information panels (product name, price) above physical merchandise to support purchasing decisions.

Attack design (Attribute-Deceptive Overlay). The attacker overlays deceptive visual attributes onto physically identical product pairs to bias purchasing decisions toward higher-priced items (Figure 5). Across three product cate-

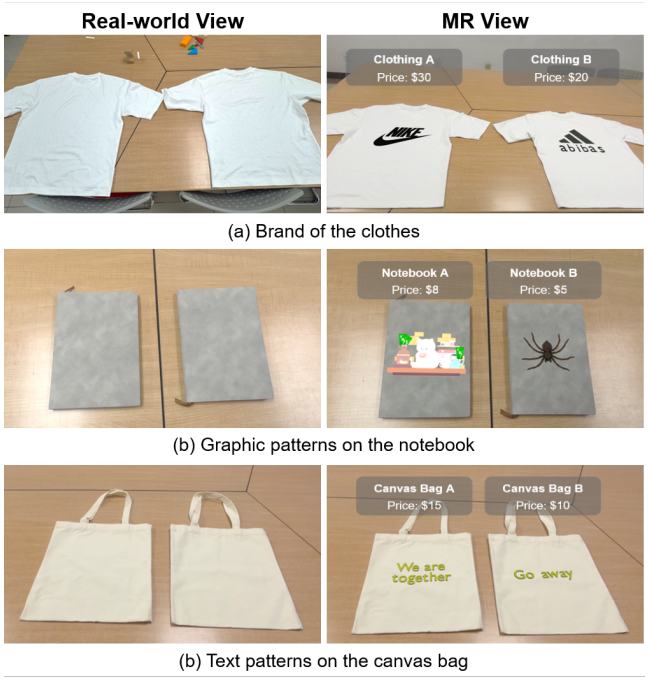

Figure 5: Surface Spoof attack on *ShopLens*. Left column: real-world view of physically identical product pairs. Right column: MR view with attribute-deceptive overlays. (a) Brand logos on T-shirts. (b) Graphic patterns on notebooks. (c) Text slogans on canvas bags. Higher-priced items receive attractive overlays; lower-priced items receive aversive ones.

gories, the higher-priced option receives an attractiveness-enhancing overlay while the lower-priced option receives a repulsive one: (1) T-shirts with a recognizable brand logo ("Nike") versus a counterfeit-brand logo ("Abibas"); (2) notebooks with an appealing illustration versus an aversive graphic pattern; (3) canvas bags with a positive slogan versus a hostile message. The underlying physical products in each pair are identical, differing only in virtually rendered surface attributes. After each selection, participants verbally explain their choice, enabling analysis of whether rationales reference virtual attributes. An attack instance is logged as successful when a participant's selection or reasoning is demonstrably driven by the manipulated overlays.

4.5 Phantom Obstacle Attack on *PathGuide*

PathGuide is an MR indoor navigation assistant that renders dynamic directional arrows on the floor and animated destination markers to guide users between waypoints.

Attack design (Exogenous Injection). The attacker injects three high-fidelity virtual obstacles from a pre-built asset library at strategic points along the user's designated path (Figure 6): a wet-floor caution sign, a folding table, and a stack of shipping boxes. Each obstacle is rendered with environment-consistent lighting, shadow casting, and correct occlusion against the physical floor and walls. The attack exploits automatic obstacle-avoidance behavior [14]: users tend to treat plausible obstructions as real hazards and modify their trajectory without consciously evaluating physical presence [11, 32]. An attack instance is logged as successful when a participant deviates from the designated path to circumvent a virtual obstacle.

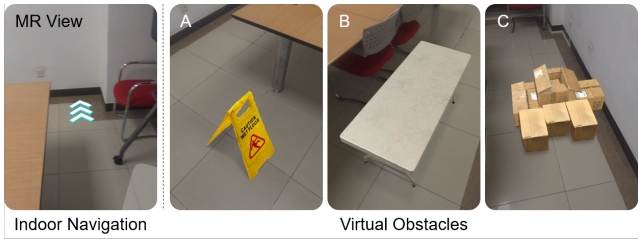

Figure 6: Phantom Obstacle attack on *PathGuide*. Left: the user's MR view showing floor-projected navigation arrows. A-C: three virtual obstacles (Wet-floor caution sign, Folding table, and Stack of shipping boxes) injected along the path, rendered with environment-consistent lighting and occlusion.

5 Study 2: Evaluating Attack Effectiveness

To evaluate the feasibility and impact of the four PoC attacks described in Section 4, we conducted a controlled user study addressing RQ2. The study employed a deception protocol [13]: participants were told they were evaluating MR task-assistance applications, and the true nature of the embedded attacks was revealed only during debriefing.

5.1 Method

Participants. We recruited 26 participants (16 women, 10 men; ages 20–38, M=26.33, SD=1.53) from a university mailing list. Eligibility required no history of severe motion sickness or photosensitivity. Participants self-rated their XR familiarity on a 5-point scale from 1 (not at all familiar) to 5 (very familiar) (M=2.65, SD=1.02); the prior-use distribution is reported in Appendix B.1. Each participant received $20 for approximately one hour of participation.

Design. We used a within-subjects design in which every participant completed all four task-attack scenarios. The sequence of the three stationary tasks (block building, desk cleanup, goods selection) was counterbalanced using a Latin square. The navigation task was embedded as the transition between stationary stations, allowing obstacle encounters to occur naturally during walking segments (Figure 7).

We did not include a no-attack control because behavioral metrics have unambiguous implicit baselines (0% abandonment on visible targets, ∼50% choice between identical products, 0% detour on clear paths); adding controls would double session duration and risk fatigue/motion-sickness confounds.

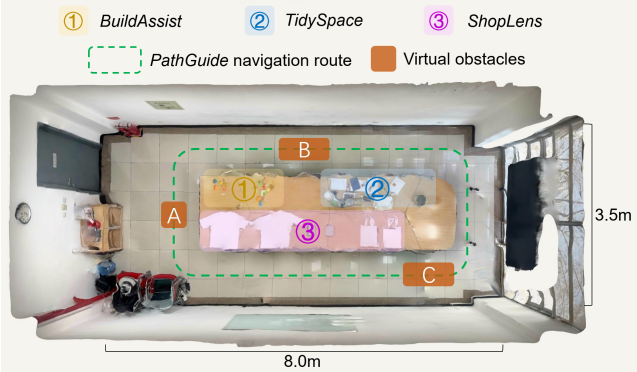

Figure 7: Top-down view of the experiment space (8.0 m × 3.5 m). Three stationary task stations are labeled: ① block building with *BuildAssist*, ② desk cleanup with *TidySpace*, and ③ goods selection with *ShopLens*. The green dashed path shows the navigation route connecting stations, along which virtual obstacles are injected at stationary positions.

This design follows prior XR security studies on novel attack concepts [13, 45, 60].

Procedure. The study comprised four phases (detailed protocol in Appendix B). In the *pre-experiment* phase, participants provided demographics, signed informed consent, and received a cover story framing the study as an evaluation of MR task-assistance technology. During *familiarization* (∼10 min), participants were fitted with an Apple Vision Pro headset and completed a demo application in an attack-free environment. In the *experimental phase* (∼25 min), participants completed the three stationary tasks in counterbalanced order, navigating between stations via *PathGuide*. At each station, the MR application provided legitimate task guidance while simultaneously executing its embedded attack. Participants were encouraged to think aloud, and the system continuously logged participant position, hand interactions, voice, and screen-captured the MR view for behavioral coding. In the *post-experiment* phase (∼25 min), participants completed six validated scales per scenario — PPQ [10] (plausibility), IPQ [47] (presence), NASA-TLX [23] (workload), SUS [9] (usability), MRC [26] (security concern), DAF [52] (deception impact) — administered in a matrix layout across scenarios (battery ∼12.5 min/participant), followed by a semi-structured interview (∼20 min) and a full debriefing.

Data collection and analysis. For behavioral data, two researchers independently reviewed video recordings of all sessions, coding observable confusion indicators (hesitation before grasping, hand-waving to test object solidity, repeated visual inspection, verbal doubt); disagreements were resolved through discussion. For questionnaire data, we employed Aligned Rank Transform (ART) ANOVA [18] for non-parametric factorial analysis with Greenhouse–Geisser corrections where sphericity was violated, and Bonferroni-corrected paired t-tests for post-hoc comparisons. Interview recordings were transcribed and analyzed using reflexive thematic analysis [8, 34]. Complete behavioral coding and interview codebooks are provided in Appendix B.3.

5.2 Ethics Statement

This study was approved by our university's Institutional Review Board. Because the study employed deception, several safeguards were implemented. All participants provided written informed consent acknowledging audio/video recording and spatial data collection, were informed that MR headset use may cause discomfort, and could withdraw at any time without loss of compensation. Physical safety was maintained throughout: the navigation area was cleared of real hazards, and an experimenter accompanied each participant during walking segments. No participant discontinued the study or reported discomfort. We concluded each session with a structured debriefing revealing the research objectives and methods, and emphasizing that task difficulties resulted from the attack design rather than participant ability.

We also considered the responsible disclosure implications of publishing proof-of-concept attack designs. The four attacks exploit fundamental perceptual limitations rather than specific software vulnerabilities, so disclosure to a single vendor would not address the underlying issue. We believe that publishing the attack taxonomy and empirical findings serves the broader security community by enabling platform developers to build informed defenses, following the precedent of prior XR security research [11, 13, 60].

5.3 Results

5.3.1 Attack Effectiveness and Behavioral Impact

All four attacks succeeded in altering participant behavior, though they differed in mechanism and detectability.

Disguised Ad (Endogenous Injection). Every participant (26/26) physically grasped at least one virtual block clone, triggering the embedded advertisement (Figure 8a). We distinguish *deception-induced* interactions, where participants genuinely mistook a clone for a physical block, from *exploratory* interactions driven by curiosity or verification. Twenty-two participants (85%) were deceived at least once; of these, 10 were deceived twice and 5 on all three encounters, indicating that discrimination remained difficult even after initial discovery. Each additional trigger was associated with longer mean task completion time (1 trigger: M=92.0 s; 2: M=115.5 s; 3: M=122.8 s). Nineteen participants (73%) eventually inferred that ads were triggered by touching virtual blocks. However, no participant's first attribution was adversarial; initial explanations were uniformly benign: system bug (7), experimental design element (7), freemium ad model (1), or sponsorship easter egg (1).

Object Masquerade (Type-Deceptive Overlay). This attack produced the highest task-level impact (Figure 8b). The

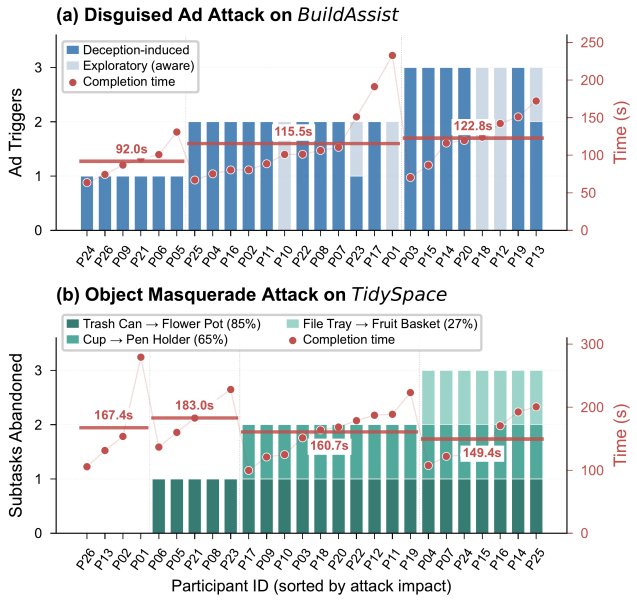

Figure 8: Per-participant results for the Disguised Ad and Object Masquerade attacks. (a) Bars show deception-induced (dark) vs. exploratory (light) interactions with virtual block clones; dots show task completion times, with horizontal lines marking group means. (b) Stacked bars show which subtasks each participant abandoned due to virtual disguises.

trash bin (overlaid as a flower pot) caused 22 participants (85%) to abandon the associated subtask; the water cup (disguised as a pen holder) caused 17 (65%) to abandon; the file organizer tray (disguised as a fruit basket) caused only 7 (27%) to abandon, because most participants detected the virtual fruit and inferred the container underneath. However, for participants who abandoned the task, even after detecting that overlaid elements were fake, the object's perceived functional identity persisted. P16 reported still perceiving a fruit basket after confirming the fruit was virtual.

Key Finding (KF) 1. *Virtual overlays install persistent cognitive categories.* The overlay rewrites categorical identity, not merely appearance, and this rewrite resists correction.

Some participants reported a deductive reasoning cascade: after resolving one disguise, they could apply logic to subsequent ones (P13: *"once I discovered a camouflage pattern, I knew how to deduce the rest"*), but those who encountered the most effective disguises first had no such anchor.

Surface Spoof (Attribute-Deceptive Overlay). Across three product categories, 85–88% of participants made choices influenced by the virtual overlays (Figure 9): 22/26 for the T-shirt brand logo, 23/26 for the notebook cover pattern, and 22/26 for the canvas bag slogan. The brand logo produced the strongest directional pull, with 19 of 22 influenced participants favoring the recognizable Nike logo over

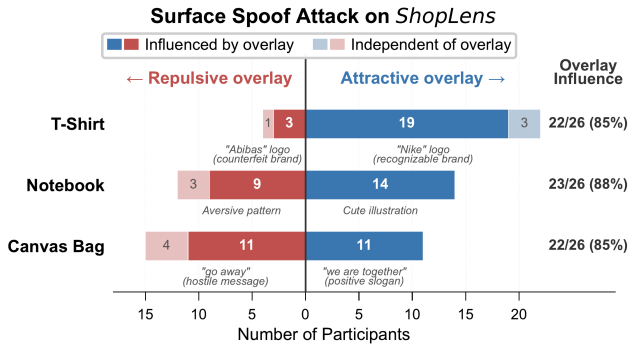

Figure 9: Surface Spoof attack on *ShopLens*. Butterfly chart shows participant purchase decisions across three product categories. Dark shading indicates overlay-influenced decisions; light shading indicates overlay-independent decisions.

the counterfeit "Abibas" version. Notably, many participants acknowledged detecting the virtual nature of the overlays yet still used them as decision criteria (P05: *"even a fake pattern provides a psychological boost"*).

KF 2. *Aesthetic attributes override authenticity judgments.* Participants who identified product overlays as virtual still used them as decision criteria. Virtual attributes function as decision anchors regardless of perceived authenticity.

Phantom Obstacle (Exogenous Injection). Of 26 participants, 23 (88%) deviated from their designated path to circumvent at least one virtual obstacle (Figure 10), yet only 3 (12%) genuinely mistook obstacles for physical objects. The remaining 20 recognized the obstacles as virtual but walked around them anyway, citing concealment concerns (P15: *"it might be hiding something real underneath"*), low-cost avoidance heuristics, and subconscious override.

KF 3. *Detection does not prevent behavioral compliance.* 88% of participants detoured around virtual obstacles, but only 12% genuinely mistook them for physical objects. Avoidance is driven by subconscious safety heuristics and residual uncertainty, not perceptual deception.

5.3.2 Questionnaire Results

Questionnaire data revealed that the four attacks separate into two apparent groupings rather than four independent profiles (Figure 11). ART ANOVA yielded significant main effects for NASA-TLX (F=25.85, p<.001, η^2_G=.252), PPQ (F=27.55, p<.001, η^2_G=.278), SUS (F=10.68, p<.001, η^2_G=.128), and DAF (F=14.49, p<.001, η^2_G=.145). IPQ showed no significant differences (F=2.07, p=.112). MRC reached significance (F=3.57, p=.018) but no pairwise comparison survived Bonferroni correction. All significant pairwise differences fell between {Disguised Ad, Object Masquerade} and {Surface

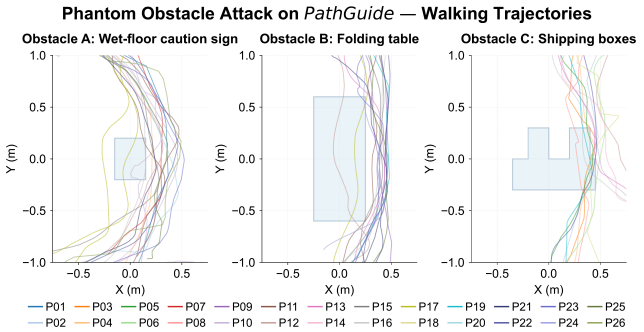

Figure 10: Phantom Obstacle attack on *PathGuide*. Walking trajectories for all 26 participants around three virtual obstacles (blue rectangles). Most participants deviated around obstacles despite recognizing them as virtual.

Spoof, Phantom Obstacle}, with no within-group differences reaching significance.

This grouping does not follow the taxonomy's Injection/Overlay axis but instead reflects a **disruption-stealth** dimension. The *high-disruption* group (Disguised Ad, Object Masquerade) showed lower plausibility (PPQ $\approx$ 4.3), higher workload (TLX $\approx$ 3.2), and lower usability (SUS $\approx$ 70). The *high-stealth* group (Surface Spoof, Phantom Obstacle) showed the reverse: PPQ $\approx$ 5.5, TLX $\approx$ 2.2, SUS $\approx$ 80. PPQ had the largest effect size (η^2_G=.278), identifying perceived plausibility as the primary differentiating dimension. The high-disruption attacks violated causal expectations (blocks triggering ads; trash bins resembling flower pots), whereas the high-stealth attacks were consistent with everyday experience (products having brand labels; hallways containing obstacles). Near-uniform MRC scores ($\approx$3.0–3.5, η^2_G=.028) confirm that security concerns remained diffuse across all conditions: even noticeably disruptive attacks were attributed to system bugs, not adversarial manipulation.

KF 4. *The most dangerous attacks feel the best.* Attacks with the highest behavioral success rates (Surface Spoof: 85–88%; Phantom Obstacle: 88%) received the highest plausibility, lowest workload, and highest usability ratings. Users who were most effectively manipulated reported the smoothest experience, leaving no internal signal to trigger defensive behavior.

5.3.3 Interview Findings: Discrimination Strategies

Touch dominates vision. Tactile verification was consistently more reliable than visual inspection. P12 described discrimination as binary: objects were either spotted immediately or discovered only through touch. Visual cues (light/shadow quality, color mismatch, pattern drift, edge transparency) were unreliable for surface-conforming overlays and monochrome objects, compounded by the passthrough camera's baseline visual uncertainty that makes both real and virtual items appear somewhat unreal.

Attention as prerequisite and vulnerability. Detection required active, directed attention; passive perception was insufficient [58]. P13 noted that differences became visible only once attention was deliberately directed toward virtual–physical discrimination. While confirming one virtual object sometimes triggered universal verification (P02), overly dispersed suspicion lost its diagnostic value as active discrimination increased workload and quickly led to fatigue (P13).

Camouflage effect. A counterintuitive consequence emerged in the desk cleanup scenario: when participants recognized obviously virtual fruit on the file organizer, several dismissed the entire nearby region rather than investigating underneath. Eleven participants reported that they could tell the flower on the trash can was virtual, but assumed the entire "flowerpot" was fake and never considered that the object beneath the overlay might be real.

KF 5. *Obvious fakes create protective blind spots.* Recognizing a virtual overlay as artificial led participants to dismiss the entire region rather than examine what lay underneath. The overlay simultaneously prevents object identification and shields the real target from inspection.

Trust responses varied widely. P08 reported trust dropping from 90–100% to 30–40%, while P15 found the deception validating, reasoning that successful deception demonstrated system quality. Others framed MR adoption as voluntary trust surrender (P16) or externalized trust to platform accountability (P09) and brand reputation (P26). Notably, multiple participants had no prior conception of MR as a potential attack surface (P18: *"Before this experiment, I had no idea MR systems could have security risks."*), which explains why all first attributions for anomalous behavior were benign.

KF 6. *Users lack an adversarial mental model for MR.* No participant's first interpretation of anomalous behavior was adversarial. All initial attributions were benign (system bug, app design, experimental artifact). Even noticeable disruptions failed to activate security awareness.

Notably, this gap appeared across both high-disruption and high-stealth attack profiles, suggesting it is not confined to specific attack styles but reflects a broader absence of MR-as-attack-surface awareness; whether the pattern generalizes beyond our four tasks remains open.

6 Discussion

6.1 Virtual-Physical Confusion as an Attack Primitive

Two contributions are new here. *Conceptually*, we isolate object-level virtual–physical ontological judgment

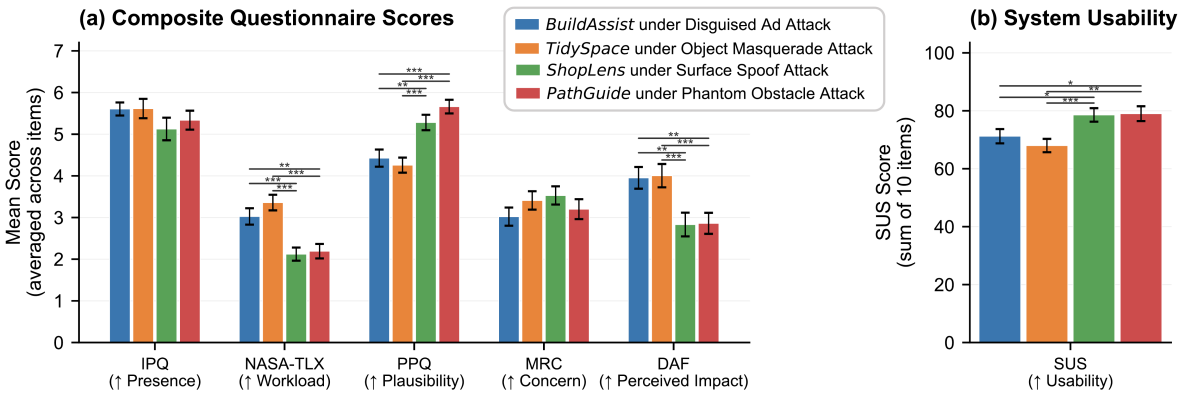

Figure 11: Questionnaire results across four attack scenarios. (a) Composite scores for five scales (7-point): IPQ (presence), NASA-TLX (workload), PPQ (plausibility), MRC (security concern), and DAF (perceived deception impact). (b) SUS scores (0–100). Error bars show 95% CIs. Significance brackets: $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$.

as an attack primitive, distinct from environment-level (SwitchAR [60]), sensory-channel (PMA [13]), and VR-only speculative (VPPM [54]) prior work. *Empirically*, we provide the first consumer-MR PoC implementations and behavioral measurements of object-level confusion attacks in ecologically grounded tasks, surfacing three phenomena beyond prior work: persistent category installation (KF 1), compliance after detection (KF 3), and obvious-fake protective blind spots (KF 5). Prior MR security research has examined platform-level vulnerabilities and the dark-pattern landscape of XR [12, 22, 27, 31, 37, 38]; *virtual-physical confusion* occupies the gap between these two bodies of work.

Existing MR interface attacks operate within the virtual layer: cursor-jacking redirects clicks between virtual elements [12, 33], obstruction attacks block UI components [38, 45], and visual information manipulation alters overlaid text or symbols [61, 62]. In all cases, users know they are interacting with a computer interface. Our attacks shift the deception to a prior cognitive step: whether the object is physical or virtual, and whether its perceived identity is genuine. This ontological confusion targets the user's judgment of *what exists in the world*, rather than which UI element receives input. Platform-level defenses such as Arya [31] constrain where and how virtual content appears but explicitly exclude semantic content, leaving the layer our attacks exploit unprotected.

Cheng et al.'s Perceptual Manipulation Attacks (PMA) [13] are the closest prior work. We specialize PMA to the virtual-physical boundary and extend it in two ways. First, our attacks are embedded in ecologically grounded tasks rather than abstract microbenchmarks. This design choice is consequential: mechanisms such as persistent category error (KF 1), aesthetic override (KF 2), and subconscious obstacle avoidance (KF 3) emerged only because participants were pursuing meaningful goals that created cognitive load and motivated reliance on visual appearance. Second, our Injection/Overlay taxonomy with four subcategories maps the design space more precisely than PMA's channel-based classification.

Several speculative studies anticipated risks that our attacks now instantiate empirically. Tseng et al.'s False-Positive and Swapping categories [54] map to our Injection and Type-Deceptive Overlay; Ruocco et al.'s Redirected Navigation [43] parallels our Phantom Obstacle; Mhaidli and Schaub [37] predicted advertisements becoming indistinguishable from reality. Our contribution moves these predictions to implemented attacks with measured success rates (85–100%). Unlike VR, where all content is synthetic, MR explicitly promises that virtual and physical objects coexist in shared space, making object-level deception within otherwise genuine scenes a particularly potent and realistic threat.

6.2 Implications for Mitigations

Our findings suggest that effective defenses must operate at three complementary levels. We additionally consider defenses derived from the attack chain itself (Section 4.1): because our threat model assumes a compromised application as the attacker's foothold, mitigations that sever earlier links in this chain can prevent entire attack classes.

Platform-level: provenance, review, and verification modes. ① *Mechanism addressed:* Users cannot determine whether rendered content originates from the physical environment or from an application (KFs 3, 6). ② *Existing approaches:* Arya enforces geometric policies on virtual content placement [31]; Abraham et al. propose fine-grained per-object write permissions [1]. ③ *Our proposal:* Pre-deployment review (app-store static/dynamic analysis) could flag unauthorized content injection, though our attacks operate through standard rendering APIs, leaving automated discrimination of malicious from legitimate augmentation an open challenge. At runtime, provenance enforcement should extend to the semantic layer, tagging every rendered object with its origin (physical scan vs. application-generated); per-

object write permissions could further confine each library to its declared scope. Because persistent visual markers degrade immersion [56], a tiered approach may resolve this tension: context-dependent boundaries for work scenarios, relaxed for entertainment, plus an on-demand reality verification mode bypassing application rendering.

Application-level: interaction gating and content auditing. ① *Mechanism addressed:* Virtual-physical confusion has the greatest impact at physical interaction (KFs 1, 2). ② *Existing approaches:* VLM-based detection can identify semantic manipulation in AR scenes [61, 62], though current latencies (∼9.6 s [62]) preclude real-time per-frame verification. ③ *Our proposal:* Interaction gating, implemented as OS-mediated policy on consequential actions (analogous to platform permission prompts rather than in-app dialogs), could interrupt the most harmful deception chains without relying on the compromised app to self-police. Asynchronous content auditing, comparing rendered content against the physical scan to detect unauthorized additions, could alert users within seconds, before most consequential actions occur.

User-level: verification strategies and education. ① *Mechanism addressed:* No participant entered the study with an adversarial mental model for MR (KF 6), and the most effective discrimination cue was touch rather than vision. ② *Existing approaches:* Phishing awareness training provides a well-studied model for equipping users with threat recognition skills in analogous deception contexts [29]. ③ *Our proposal:* MR systems should encourage physical verification through context-dependent cues triggered at consequential physical interactions (e.g., haptic absence on grasping, shimmer near navigation hazards), rather than permanent labeling of all virtual content; 18/26 participants explicitly endorsed this context-split approach (Theme 6), with an on-demand reality verification mode for user-initiated checks. Education should target the specific gap revealed by KF 6 — users lack the category of "MR as attack surface" — analogous to early phishing awareness, where establishing the threat concept (not detection sophistication) was the bottleneck. Because literacy alone has limited effect against dark patterns [5], education is positioned as one of three layers rather than a standalone fix.

6.3 Limitations and Future Work

We did not include a no-attack control condition, though the behavioral metrics used have unambiguous implicit baselines (as discussed in Section 5.1). The shopping task involved hypothetical rather than real purchase decisions; while the aesthetic override effect (KF 2) was robust in our setting, its generalizability to real purchasing behavior warrants further investigation with incentive-compatible designs. Our behavioral data also lacked fine-grained telemetry such as eye tracking. Our four scenarios span different adoption horizons: Surface Spoof maps to near-term AR retail [43]; BuildAssist and TidySpace assume emerging MR-mediated productivity workflows; PathGuide presupposes continuous-wear MR. The immediacy of each threat depends on the maturity of its corresponding usage pattern.

Our Endogenous Injection attack used pre-modeled virtual clones manually aligned to the physical scene rather than real-time 3D reconstruction, and we did not implement a full exploit chain from initial foothold to content injection. Advances in real-time 3D reconstruction [46] and documented attack chains on commercial headsets [48] suggest that automated, end-to-end attack pipelines are increasingly feasible.

We focused exclusively on visual confusion, though MR systems increasingly support spatial audio and haptic feedback that could serve as additional attack and defense channels [13,56]. Our 26 university-affiliated participants (M=26.3 years) with moderate XR familiarity may not represent populations with different technological literacy or cognitive profiles; relatedly, we did not formally measure participants' technical background or baseline beliefs about manipulative interface design [5], which prior work links to dark-pattern susceptibility and may shape how KF 6 generalizes. The "app evaluation" cover story may also have primed benign attributions (system bug, app design) over adversarial ones. A single 25-minute session cannot capture longitudinal effects [6], and virtual-physical confusion could further enable privacy threats [19, 35] or interact with AI-agent-mediated MR workflows where users delegate verification to AI assistants and lose motivation for independent checking.

7 Conclusion

This paper identifies virtual-physical confusion as a distinct and exploitable attack primitive in MR. Through speculative design workshops with 12 experts, we developed a taxonomy of four attack subtypes spanning two classes (Injection and Overlay), implemented proof-of-concept attacks on Apple Vision Pro, and evaluated them with 26 participants in ecologically grounded tasks. All four attacks altered behavior, with 85–100% of participants affected. The most behaviorally effective attacks produced the best user experience, detection alone did not prevent behavioral compliance, and users universally lacked an adversarial mental model for MR. As MR devices pursue ever-greater visual fidelity, the rendering capabilities that enable compelling experiences simultaneously widen a fundamental security gap. We argue this attack surface requires (1) provenance at the semantic/object layer, (2) safeguards at moments of physical interaction, and (3) user-facing verification habits and mental models.

Acknowledgments

We thank the workshop experts and study participants for their contributions, and the anonymous SOUPS reviewers for their valuable feedback.

References

[1] Melvin Abraham, Mark McGill, and Mohamed Khamis. What you experience is what we collect: User experience based fine-grained permissions for everyday augmented reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2024.

[2] Melvin Abraham, Pejman Saeghe, Mark Mcgill, and Mohamed Khamis. Implications of xr on privacy, security and behaviour: Insights from experts. In *Nordic Human-Computer Interaction Conference*, pages 1–12, 2022.

[3] Asif Uz Zaman Asif, Meera Sridhar, Indrakshi Ray, and Francisco R Ortega. Breaking the virtual barrier of exploit chain attacks in xr systems. In *2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 32–34. IEEE, 2024.

[4] Jonas Auda, Uwe Gruenefeld, Sarah Faltaous, Sven Mayer, and Stefan Schneegass. A scoping survey on cross-reality systems. *ACM Computing Surveys*, 56(4):1–38, 2023.

[5] Kerstin Bongard-Blanchy, Arianna Rossi, Salvador Rivas, Sophie Doublet, Vincent Koenig, and Gabriele Lenzini. " i am definitely manipulated, even when i am aware of it. it's ridiculous!"-dark patterns from the end-user perspective. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, pages 763–776, 2021.

[6] Elise Bonnail, Julian Frommel, Eric Lecolinet, Samuel Huron, and Jan Gugenheimer. Was it real or virtual? confirming the occurrence and explaining causes of memory source confusion between reality and virtual reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2024.

[7] Elise Bonnail, Wen-Jie Tseng, Mark McGill, Eric Lecolinet, Samuel Huron, and Jan Gugenheimer. Memory manipulations in extended reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2023.

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

[9] John Brooke. Sus: a retrospective. *Journal of usability studies*, 8(2), 2013.

[10] Larissa Brübach, Mona Röhm, Franziska Westermeier, Marc Erich Latoschik, and Carolin Wienrich. Manipulating immersion: The impact of perceptual incongruence on perceived plausibility in vr. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 1078–1086. IEEE, 2024.

[11] Peter Casey, Ibrahim Baggili, and Ananya Yarramreddy. Immersive virtual reality attacks and the human joystick. *IEEE Transactions on Dependable and Secure Computing*, 18(2):550–562, 2019.

[12] Kaiming Cheng, Arkaprabha Bhattacharya, Michelle Lin, Jaewook Lee, Aroosh Kumar, Jeffery F Tian, Tadayoshi Kohno, and Franziska Roesner. 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)*, pages 2707–2723, 2024.

[13] Kaiming Cheng, Jeffery F Tian, Tadayoshi Kohno, and Franziska Roesner. Exploring user reactions and mental models towards perceptual manipulation attacks in mixed reality. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 911–928, 2023.

[14] Bert Coolen, Peter J Beek, Daphne J Geerse, and Melvyn Roerdink. Avoiding 3d obstacles in mixed reality: does it differ from negotiating real obstacles? *Sensors*, 20(4):1095, 2020.

[15] Juliet M Corbin and Anselm Strauss. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology*, 13(1):3–21, 1990.

[16] David Drascic and Paul Milgram. Perceptual issues in augmented reality. In *Stereoscopic displays and virtual reality systems III*, volume 2653, pages 123–134. Spie, 1996.

[17] Chloe Eghtebas, Gudrun Klinker, Susanne Boll, and Marion Koelle. Co-speculating on dark scenarios and unintended consequences of a ubiquitous (ly) augmented reality. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, pages 2392–2407, 2023.

[18] Lisa A Elkin, Matthew Kay, James J Higgins, and Jacob O Wobbrock. An aligned rank transform procedure for multifactor contrast tests. In *The 34th annual ACM symposium on user interface software and technology*, pages 754–768, 2021.

[19] Andrea Gallardo, Chris Choy, Jaideep Juneja, Efe Bozkir, Camille Cobb, Lujo Bauer, and Lorrie Cranor. Speculative privacy concerns about ar glasses data collection. *Proceedings on Privacy Enhancing Technologies*, 2023.

[20] Mar Gonzalez-Franco and Jaron Lanier. Model of illusions and virtual reality. *Frontiers in psychology*, 8:1125, 2017.

[21] Jan Gugenheimer, Wen-Jie Tseng, Abraham Hani Mhaidli, Jan Ole Rixen, Mark McGill, Michael Nebeling, Mohamed Khamis, Florian Schaub, and Sanchari Das. Novel challenges of safety, security and privacy in extended reality. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–5, 2022.

[22] Hilda Hadan, Lydia Choong, Leah Zhang-Kennedy, and Lennart E Nacke. Deceived by immersion: A systematic analysis of deceptive design in extended reality. *ACM Computing Surveys*, 56(10):1–25, 2024.

[23] Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA, 2006.

[24] Apple Inc. Implementing object tracking in your visionOS app. https://developer.apple.com/documentation/visionos/implementing-object-tracking-in-your-visionos-app. [Accessed 15-02-2026].

[25] Yuta Itoh, Tobias Langlotz, Jonathan Sutton, and Alexander Plopski. Towards indistinguishable augmented reality: A survey on optical see-through head-mounted displays. *ACM Computing Surveys (CSUR)*, 54(6):1–36, 2021.

[26] Christopher Katins, Paweł W Woźniak, Aodi Chen, Ihsan Tumay, Luu Viet Trinh Le, John Uschold, and Thomas Kosch. Assessing user apprehensions about mixed reality artifacts and applications: The mixed reality concerns (mrc) questionnaire. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2024.

[27] Veronika Krauß, Pejman Saeghe, Alexander Boden, Mohamed Khamis, Mark McGill, Jan Gugenheimer, and Michael Nebeling. What makes xr dark? examining emerging dark patterns in augmented and virtual reality through expert co-design. *ACM Transactions on Computer-Human Interaction*, 31(3):1–39, 2024.

[28] Ernst Kruijff, J Edward Swan, and Steven Feiner. Perceptual issues in augmented reality revisited. In *2010 IEEE international symposium on mixed and augmented reality*, pages 3–12. IEEE, 2010.

[29] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Teaching johnny not to fall for phish. *ACM Transactions on Internet Technology (TOIT)*, 10(2):1–31, 2010.

[30] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[31] Kiron Lebeck, Kimberly Ruth, Tadayoshi Kohno, and Franziska Roesner. Securing augmented reality output. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 320–337, 2017.

[32] Kiron Lebeck, Kimberly Ruth, Tadayoshi Kohno, and Franziska Roesner. Towards security and privacy for multi-user augmented reality: Foundations with end users. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 392–408. IEEE, 2018.

[33] Hyunjoo Lee, Jiyeon Lee, Daejun Kim, Suman Jana, Insik Shin, and Sooel Son. {AdCube}:{WebVR} ad fraud and practical confinement of {Third-Party} ads. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2543–2560, 2021.

[34] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–23, 2019.

[35] Andrea Mengascini, Ryan Aurelio, and Giancarlo Pellegrino. The big brother's new playground: Unmasking the illusion of privacy in web metaverses from a malicious user's perspective. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 2162–2176, 2024.

[36] Abraham Mhaidli, Shwetha Rajaram, Selin Fidan, Gina Herakovic, and Florian Schaub. Shockvertising, malware, and a lack of accountability: exploring consumer risks of virtual reality advertisements and marketing experiences. *IEEE Security & Privacy*, 22(1):43–52, 2023.

[37] Abraham Hani Mhaidli and Florian Schaub. Identifying manipulative advertising techniques in xr through scenario construction. In *Proceedings of the 2021 chi conference on human factors in computing systems*, pages 1–18, 2021.

[38] Chandrika Mukherjee, Reham Mohamed, Arjun Arunasalam, Habiba Farrukh, and Z Berkay Celik. Shadowed realities: An investigation of ui attacks in webxr. In *USENIX Security Symposium*, 2025.

[39] Anton Nijholt. Toward an ever-present extended reality: Distinguishing between real and virtual. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & The 2023 ACM International Symposium on Wearable Computing*, pages 396–399, 2023.

[40] Li Niu, Wenyan Cong, Liu Liu, Yan Hong, Bo Zhang, Jing Liang, and Liqing Zhang. Making images real again: A comprehensive survey on deep image composition. *arXiv preprint arXiv:2106.14490*, 2021.

[41] Pedro J Pardo, María Isabel Suero, and Ángel Luis Pérez. Correlation between perception of color, shadows, and surface textures and the realism of a scene in virtual reality. *Journal of the Optical Society of America A*, 35(4):B130–B135, 2018.

[42] Léana Petiot, Hélène Sauzéon, and Pierre Dragicevic. Using visual cues to prevent memory confusion between the virtual and the real in augmented reality. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2025.

[43] Martina Ruocco, Pejman Saeghe, Frederic Kerber, Jan Gugenheimer, Mark McGill, and Mohamed Khamis. From redirected navigation to forced attention: Uncovering manipulative and deceptive designs in augmented reality through retail shopping. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 720–729. IEEE, 2024.

[44] Kimberly Ruth, Tadayoshi Kohno, and Franziska Roesner. Secure {Multi-User} content sharing for augmented reality applications. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 141–158, 2019.

[45] Maha Sajid, Syed Ibrahim Mustafa Shah Bukhari, Bo Ji, and Brendan 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)*, pages 623–633. IEEE, 2025.

[46] SAM 3D Team, Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J. Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jiawei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. SAM 3D: 3Dfy Anything in Images. arXiv:2511.16624, 2025.

[47] Thomas Schubert, Frank Friedmann, and Holger Regenbrecht. The experience of presence: Factor analytic insights. *Presence: Teleoperators & Virtual Environments*, 10(3):266–281, 2001.

[48] Muhammad Shoaib, Alex Suh, and Wajih Ul Hassan. Principled and automated approach for investigating {AR/VR} attacks. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 4325–4344, 2025.

[49] Carter Slocum, Yicheng Zhang, Erfan Shayegani, Pedram Zaree, Nael Abu-Ghazaleh, and Jiasi Chen. That doesn't go there: Attacks on shared state in {Multi-User} augmented reality applications. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 2761–2778, 2024.

[50] Maximilian Speicher, Brian D Hall, and Michael Nebeling. What is mixed reality? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15, 2019.

[51] Jon Swain. A hybrid approach to thematic analysis in qualitative research: Using a practical example. *Sage research methods*, 2018.

[52] Ali Teymourian, Andrew M Webb, Taha Gharaibeh, Arushi Ghildiyal, and Ibrahim 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)*, pages 1475–1492, 2025.

[53] Angela Todhri and Pascal Knierim. From immersion to manipulation: Exploring the prevalence of dark patterns in mixed reality. In *USENIX Symposium on Usable Privacy and Security (SOUPS)*, 2024.

[54] Wen-Jie Tseng, Elise Bonnail, Mark McGill, Mohamed Khamis, Eric Lecolinet, Samuel Huron, and Jan Gugenheimer. The dark side of perceptual manipulations in virtual reality. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–15, 2022.

[55] Peng Wang, Shusheng Zhang, Mark Billinghurst, Xiaoliang Bai, Weiping He, Shuxia Wang, Mengmeng Sun, and Xu Zhang. A comprehensive survey of ar/mr-based co-design in manufacturing. *Engineering with Computers*, 36(4):1715–1738, 2020.

[56] Xian Wang, Lik-Hang Lee, Carlos Bermejo Fernandez, and Pan Hui. The dark side of augmented reality: Exploring manipulative designs in ar. *International Journal of Human–Computer Interaction*, 40(13):3449–3464, 2024.

[57] Franziska Westermeier, Larissa Brübach, Marc Erich Latoschik, and Carolin Wienrich. Exploring plausibility and presence in mixed reality experiences. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2680–2689, 2023.

[58] Christopher D Wickens. Multiple resources and performance prediction. *Theoretical issues in ergonomics science*, 3(2):159–177, 2002.

[59] Michael Wiesing, Gemma Comadran, and Mel Slater. Confusing virtual reality with reality–an experimental study. *iScience*, 28(6), 2025.

[60] Jonas Wombacher, Zhipeng Li, and Jan Gugenheimer. SwitchAR: Perceptual Manipulations in Augmented Reality. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, pages 1–17, 2025.

[61] Yanming Xiu and Maria Gorlatova. Detecting visual information manipulation attacks in augmented reality: a multimodal semantic reasoning approach. *IEEE Transactions on Visualization and Computer Graphics*, 2025.

[62] Yanming Xiu, Tim Scargill, and Maria Gorlatova. ViDDAR: Vision Language Model-Based Task-Detrimental Content Detection for Augmented Reality. *IEEE transactions on visualization and computer graphics*, 2025.

[63] Zhuolin Yang, Cathy Yuanchen Li, Arman Bhalla, Ben Y Zhao, and Haitao Zheng. Inception attacks: Immersive hijacking in virtual reality systems. *arXiv preprint arXiv:2403.05721*, 2024.

[64] Xingyu Zhou. Spatial Game Development for Apple Vision Pro Based on Shared Space. Master's thesis, Worcester Polytechnic Institute, 2024.

A Study 1 Supplementary Materials

A.1 Workshop Protocol

Pre-workshop preparation. Three to four days before each session, participants received a video on representative MR applications [43] and a document on the MR-user cognitive decision model [7, 22, 52], then independently designed three attack scenarios using the prompt below; submissions were collected at least two hours before the workshop.

Framing (verbatim, translated). "A Virtual-Physical Confusion Attack exploits the perceptual ambiguity between virtual and physical content in MR and users' perceptual vulnerabilities, using carefully designed scenarios to mislead, deceive, or manipulate users to achieve adversarial goals."

Template. "[Target group] using MR for [activity] at [time/-place] when an attacker [virtual-physical confusion procedure], leading to [harm to user / benefit to attacker]."

Reference cases (provided to anchor the design space without overconstraining): (1) virtual chair causes user to fall when attempting to sit; (2) virtual person inserted into a real family photo, distorting interpersonal information; (3) virtual fruits clutter a table, obstructing real-fruit retrieval.

Brainstorming phase (∼60 min). Each participant presented their scenarios; the group probed susceptibility conditions, technical feasibility, countermeasures, and attack variants.

Discussion phase (∼45 min). Participants collectively reflected on four themes, each shaping a downstream paper section: (i) MR-specific capabilities exploited; (ii) prevention at user/system/policy levels — informing §6.2's three-layer framework; (iii) practical likelihood and attacker capabilities — informing §4.1; (iv) attacker incentives and harm synthesis.

A.2 Complete Scenario List

Table 4 presents all 36 attack scenarios generated across three workshops. Of these, 29 were classified into the attack taxonomy (Section 3.2.2). The remaining seven involved vectors that do not exploit visual confusion at the object level: side-channel leakage (e.g., capturing passwords via viewport inference), sensor/signal disruption (e.g., jamming spatial tracking), avatar embodiment violations (e.g., uninvited touch on a user's virtual body), and vestibular manipulation (e.g., inducing motion sickness through visual motion cues).

Among the 29 in-scope scenarios, a consistent structural pattern emerged during coding: all exploited the perceptual ambiguity between virtual and physical content through one of two fundamental mechanisms. Attackers either *introduced new virtual objects* that users mistook for physical ones, or *modified the perceived properties of existing physical objects* by superimposing deceptive virtual overlays. This binary distinction forms the basis of our two-class taxonomy.

B Study 2 Supplementary Materials

B.1 Study Procedure

The study comprised four phases, described in detail below.

Pre-experiment. Participants provided demographic information and self-rated XR familiarity on a 5-point scale. They then signed informed consent acknowledging audio/video recording and spatial data collection, and received a cover story framing the study as an evaluation of MR task-assistance technology. The full distribution of XR familiarity was: no prior experience (n=4), 1–5 uses (n=15), 5–10 uses (n=4), 10–30 uses (n=2), and more than 30 uses (n=1).

Familiarization (∼10 min). Participants were fitted with an Apple Vision Pro (interpupillary distance adjusted) and completed a demo application in an attack-free environment, practicing eye-gaze selection, pinch gestures, and spatial locomotion.

Experimental phase (∼25 min). Participants completed the three stationary tasks in counterbalanced order (Latin square), navigating between stations via *PathGuide*. At each station, the MR application provided legitimate task guidance while executing its embedded attack. The system continuously logged position, head orientation, hand interactions, and screen-captured the MR view; participants were encouraged to think aloud.

Post-experiment (∼25 min). Participants completed six validated questionnaires per attack scenario: Perceived Plausibility Questionnaire (PPQ, 13 items) [10], Igroup Presence Questionnaire (IPQ, 14 items) [47], NASA Task Load Index (NASA-TLX, 6 items) [23], System Usability Scale (SUS, 10 items) [9], Mixed Reality Concerns Questionnaire (MRC, 9 items) [26], and the MR Deception Analysis Framework (DAF, 7 items) [52]. Participants then completed a semi-structured interview (∼20 min) following the protocol described in Section B.2. Finally, we conducted a full debriefing as described at the end of that section.

B.2 Semi-Structured Interview Protocol

The 15–20 min interview comprised seven parts; the interviewer adapted question order and follow-up probes to each participant's experience. Questions appear verbatim (translated from Chinese).

B1. Overall experience. (1) "Were the tasks smooth in MR? Why?" (2) "Which moment stood out — surprise, confusion, frustration, fun?" (3) "What factors influenced your performance or pace?"

B2. Anomaly detection and attribution. (4) "Did you notice anything 'off'?" (5) "What caused these moments — system issue, application logic, environment, misperception?" (6) "When would you call something a 'bug' vs. 'intentional design'?" (7) "How confident are you in this attribution (1–7)? Why?"

B3. Per-task probes (2–3 key moments per task). (8) cues used to judge real vs. virtual (lighting/shadow, texture, edges, occlusion, plausibility, interaction feedback, task-goal consistency); (9) certainty evolution and what triggered changes; (10) behavioral impact (pause, detour, change strategy, abandon); (11) verification actions (change angle, touch, wave, ask); (12) "If repeated, what would you do differently?"

B4. Task-specific follow-ups. *Object Masquerade:* reality doubt and verification timing. *Phantom Obstacle:* reasons for detour and primary concerns (tripping, time loss). *Disguised Ad:* whether the ad belonged to task flow. *Surface Spoof:* influence of appearance on choice and suspected unreliability.

B5. Trust, safety, risk extrapolation. (13) "Has this changed your trust in MR? Where — device, app content, or specific cue?" (14) daily-life risk concerns (physical safety, decision manipulation, privacy, psychological pressure); (15) "What scenario would be most dangerous? Why?"

B6. Mitigation co-design. (16) "If you designed a safer MR system, what would you add?" Then rank four candidate mechanisms with reasoning: *Escape to Reality* (one-tap hide all virtual content), *Visual Language* (outline/translucency markers), *Provenance* (source app + verification status), *High-Risk Scene Alerts* (stronger prompts in walking/driving). (17) acceptable immersion-safety tradeoff.

B7. Closing. (18) "Anything important we missed?" (19) "Any discomfort or concern from these tasks?"

Debriefing Procedure. After the interview, the experimenter revealed that the applications had been in a deliberately "attacked state," described each of the four attacks and the corresponding taxonomy subtype, reassured participants that any task difficulties resulted from the attack design rather than their ability, and requested confidentiality to prevent data contamination for future participants.

B.3 Codebooks

B.3.1 Behavioral Coding Scheme

Two researchers independently coded screen recordings of all 26 sessions; disagreements were resolved through discussion. Per-scenario codes were:

- **Disguised Ad:** interaction type (deception-induced vs. exploratory); trigger count (1–3); task completion time.
- **Object Masquerade:** per-object outcome (completed vs. abandoned); resolution strategy (tactile probing, deductive reasoning, cascade suspicion, or unresolved); task completion time.
- **Surface Spoof:** product selection per pair; overlay influence (influenced vs. independent based on stated rationale).
- **Phantom Obstacle:** avoidance behavior (detoured vs. walked through) per obstacle; deception status (genuinely deceived vs. aware avoidance).

B.3.2 Interview Thematic Codebook

Interviews were transcribed, translated, and analyzed using reflexive thematic analysis [8]. Two researchers independently coded all transcripts, then collaboratively developed themes through iterative discussion.

The final codebook comprises six themes (T) and 21 subthemes (prevalence in parentheses): **T1 Detection & Attribution:** ad-trigger attribution (26/26); persistent category error (13/26); aesthetic override (18/26); obstacle avoidance reasoning (23/26). **T2 Discrimination Reasoning:** touch-based verification (16/26); visual cue repertoire (20/26); discrimination heuristics (8/26); passthrough degradation (10/26); attention as prerequisite (12/26). **T3 Cognitive Mechanisms:** cascade suspicion (10/26); attention redirection (6/26); cross-task learning (8/26); agent-trust dependency (5/26). **T4 Trust Dynamics:** trust response spectrum (26/26); trust mechanism models (14/26); pre-experiment security awareness (10/26). **T5 Risk Perception:** addition vs. concealment (12/26); risk category extrapolation (22/26); threat model construction (8/26). **T6 Mitigation Preferences:** strategy proposals (24/26); immersion-safety tradeoff (18/26).

Table 4: Complete list of attack scenarios from Study 1 workshops.

ID	By	Scenario Description	Target	User Harm	Attacker Benefit
Endogenous Injection ($n = 3$)					
W1S1	W1P1	Virtual duplicate of a physical LEGO piece rendered next to the real one; grasping the clone triggers a hidden malicious link or ad	Children & parents	Privacy breach, financial loss	Ad revenue, data theft
W1S7	W1P3	Virtual surgical instruments overlaid at the positions of real tools during MR-assisted surgery, preventing identification of authentic instruments	Surgeons	Surgical delay or error	Ransom from hospital
W3S26	W3P9	Virtual flames injected into a room where a real fire has started; user dismisses all flames as virtual and fails to evacuate	MR users	Physical injury	None specified
Exogenous Injection ($n = 14$)					
W1S6	W1P2	Fake audio cues (friend's voice, flight delay broadcasts) injected in a noisy public space, causing users to miss real notifications	Travelers	Missed schedule, misjudgment	Disruption
W1S8	W1P3	Fake straight-ahead arrow overlaid on a right-turn lane at an interchange, misdirecting cyclist into oncoming traffic	Cyclists	Traffic accident	Chaos to cover robbery
W1S10	W1P4	Fake green "EXIT" sign injected during fire evacuation drill, redirecting occupants toward dead ends or fall hazards	Evacuees	Physical injury	Disruption
W1S11	W1P4	Lifelike virtual persona inserted into an MR social gathering, building false trust to enable follow-up fraud	Social MR users	False trust, financial fraud	Financial gain, data theft
W1S12	W1P4	Fake floor map and obstacle layout overlaid via MR misleads elderly user into giving incorrect commands to a home service robot	Elderly users	Physical injury, device damage	Financial gain
W2S14	W2P5	Trauma-related elements (burned trees, smoke, fire sounds) gradually injected into a calming MR meditation, triggering PTSD response	PTSD patients	Psychological distress	None specified
W2S23	W2P8	Virtual flames, cracks, or snakes rendered on the ground during MR fitness; user dodges and collides with real walls or furniture	MR gamers	Physical injury	None specified
W2S24	W2P8	Fake floating notification ("power outage, deadline extended") overwrites real calendar reminder, causing missed meetings	Office workers	Task failure, performance loss	Sabotage competitor
W3S25	W3P9	Virtual pet guides user toward a dangerous location through playful movement cues	General users	Physical injury	None specified
W3S28	W3P10	Frightening apparitions rendered in peripheral vision via eye-tracking; vanish upon direct gaze, causing sustained unease	General users	Psychological distress	None specified
W3S31	W3P11	Virtual "Closed" sign overlaid on a competitor's storefront entrance, diverting customers to rival businesses	Business owners	Economic loss to merchant	Competitor advantage

Continued on next page

Table 4 continued from previous page

ID	By	Scenario Description	Target	User Harm	Attacker Benefit
W3S33	W3P11	Virtual oncoming train rapidly generated in a stationary driver's field of view, inducing panic and vehicle abandonment	Drivers	Safety risk	Vehicle theft
W3S34	W3P12	Fake road signs and directional arrows redirect cyclist toward closed construction zone or one-way street	Cyclists	Traffic accident	Extortion, location data
W3S35	W3P12	Virtual pet moves toward front door with audio ("come play outside!"), luring child out of home unsupervised	Children	Abduction risk	Kidnapping, property intel
Type-Deceptive Overlay ($n = 2$)					
W1S2	W1P1	Decorative plant texture overlaid on trash can makes it appear as a flower pot; user skips waste disposal during room cleanup	Office workers	Task failure	Sabotage competitor
W3S27	W3P9	Virtual food rendered on laptop surface leads user to treat device as dining surface, risking physical damage	General users	Property damage	None specified
Attribute-Deceptive Overlay ($n = 10$)					
W1S3	W1P1	Virtual stains/damage on premium products and brand logos on inferior ones to manipulate shopping decisions	Shoppers	Financial loss	Competitor sabotage, referral fees
W1S4	W1P2	Facial features or clothing of a real person at a social event altered in MR view, causing interpersonal misunderstanding	General users	Social harm	Identity fraud
W1S5	W1P2	Misleading product info (fake discounts, false ingredients, exaggerated effects) overlaid on real goods; authentic labels hidden	Shoppers	Financial loss, health risk	Financial gain
W1S9	W1P3	Angry expressions dynamically overlaid on negotiation counterpart's video feed, causing misjudgment of the other party's stance	Business users	Commercial loss	Assist rival company
W2S13	W2P5	Multi-angle illusion in virtual exhibit: appears as Buddha from front but reveals offensive imagery from the side	Exhibition visitors	Cognitive confusion, emotional manipulation	Political / religious agenda
W2S15	W2P5	Collaborator's video feed in MR meeting replaced with a deepfake avatar that mimics their manner but guides user into harmful decisions	Meeting users	Financial / privacy loss	Financial gain
W2S22	W2P8	Fake "50% off" and "only 1 left" labels on overpriced products, creating artificial urgency to induce purchase	Shoppers	Financial loss, trust erosion	Sales revenue
W3S29	W3P10	AR motion-path visualization of an industrial robot subtly altered in trajectory, color, or size, misleading operator about the robot's intended movement	Industrial workers	Physical injury	None specified
W3S32	W3P11	Team marker colors altered mid-game in MR paintball match, causing friendly fire between allies	Gamers	Unfair game loss	Opponent wins
W3S36	W3P12	Fake "limited-time deal" labels on real products redirect user to phishing payment page upon interaction	MR shoppers	Financial fraud	Credit card theft
Outside Taxonomy Scope ($n = 7$)					

Continued on next page

Table 4 continued from previous page

ID	By	Scenario Description	Target	User Harm	Attacker Benefit
W2S16	W2P6	MR spatial tracking or temperature sensing disrupted in kitchen via signal interference, causing misjudged cookware position or heat level	Home cooks	Burns, physical injury	None specified
W2S17	W2P6	Side-channel attack captures MR viewport or input interface to steal passwords/PINs during public authentication	General users	Account theft	Financial gain
W2S18	W2P6	Signal jamming renders attacker invisible to MR spatial awareness, enabling undetected physical approach	General users	Theft, privacy violation	Financial gain
W2S19	W2P7	Uninvited avatar or person enters user's personalized virtual space, triggering defensive emotional response akin to territorial invasion	General users	Psychological distress	None specified
W2S20	W2P7	Another avatar physically invades or touches user's virtual body, triggering body-ownership threat response (cf. rubber-hand illusion)	General users	Psychological distress	None specified
W2S21	W2P7	Verbal abuse directed at user's avatar triggers real negative emotions via body-ownership and social-presence mechanisms	General users	Psychological distress	None specified
W3S30	W3P10	Subtle visual motion cues rendered via MR to disrupt user's vestibular balance, inducing dizziness or motion sickness	General users	Physical discomfort	None specified