On the Feasibility of Using MultiModal LLMs to Execute AR Social Engineering Attacks

Ting Bi[1], Chenghang Ye[2*], Zheyu Yang[2*], Ziyi Zhou[2*], Cui Tang[2*], Zui Tao[1*], Jun Zhang[1*], Kailong Wang[1], Liting Zhou[3], Yang Yang[2†], Tianlong Yu[2†]

[1]Huazhong University of Science and Technology, Wuhan, China
[2]Hubei University, Wuhan, China
[3]Dublin City University, Dublin, Ireland
ting.bi@ieee.org, (chenghangye, yangzheyu, 202421121013087, cuitang)@stu.hubu.edu.cn, (tzzz1, junzhang02, wangkl)@hust.edu.cn, liting.zhou@dcu.ie, yangyang@hubu.edu.cn, tommyyu21@163.com

Abstract

Augmented Reality (AR) and Multimodal Large Language Models (LLMs) are rapidly evolving, providing unprecedented capabilities for human-computer interaction. However, their integration introduces a new attack surface for Social Engineering (SE). In this paper, we systematically investigate the feasibility of orchestrating AR-driven Social Engineering attacks using Multimodal LLM for the first time, via our proposed SEAR framework, which operates through three key phases: (1) AR-based social context synthesis, which fuses Multimodal inputs (visual, auditory and environmental cues); (2) role-based Multimodal RAG (Retrieval-Augmented Generation), which dynamically retrieves and integrates social context; and (3) ReInteract social engineering agents, which execute adaptive multiphase attack strategies through inference interaction loops. To verify SEAR, we conducted an IRB-approved study with 60 participants and build a novel dataset of 180 annotated conversations in different social scenarios (e.g., coffee shops, networking events). Our results show that SEAR is highly effective at eliciting high-risk behaviors (e.g., 93.3% of participants susceptible to email phishing). The framework was particularly effective in building trust, with 85% of targets willing to accept an attacker's call after an interaction. Also, we identified notable limitations such as authenticity gaps. This work provides proof-of-concept for AR-LLM driven social engineering attacks and insights for developing defenses against next-generation AR/LLM-based SE threats.

Code — https://github.com/2192537130/searsystem.git

Introduction

The rapid development of Augmented Reality (AR) and Large Language Models (LLMs) is revolutionizing human-computer interaction, enabling immersive experiences that blend digital overlays with real-world environments. AR systems, equipped with Multimodal sensors like RGB-D cameras and microphones, capture rich contextual data (e.g., facial or vocal information), while LLMs analyze and generate human-like dialogue with remarkable adaptability. While this synergy enables transformative applications, it also introduces unprecedented risks: the integration of AR's real-time environmental perception and LLMs' adaptive reasoning creates a potent vector for next-generation social engineering attacks (Choo 2025).

Traditional social engineering techniques, such as phishing emails or identity theft (Ho et al. 2019; Bilge et al. 2009; Roy et al. 2024; Timko, Castillo, and Rahman 2025), rely on static deception strategies (Burda, Allodi, and Zannone 2024; Vadrevu and Perdisci 2019; Yang et al. 2023; Ulqinaku et al. 2021). In contrast, the fusion of AR's environmental perception and LLMs' generative capabilities will introduce a potential paradigm shift- allowing the attackers to craft highly personalized and adaptive attacks. For instance, AR sensors can infer a victim's emotional status during a conversation (Xu, Hou, and Jiang 2025), while LLMs can generate strategical dialogue (e.g., gradual trust-building) to exploit the reduced vigilance.

Despite the enthusiasm for AR-LLM social applications (Yang et al. 2025; Jansen and Fischbach 2020; Fuste and Schmandt 2017; Hirskyj-Douglas et al. 2020) and the growing awareness of AR privacy risks (Chen et al. 2018; Lehman et al. 2022; Deng, Zhai, and Yang 2023) and LLM-enabled phishing (Falade 2023a), no prior work systematically examines their potential for orchestrated Social Engineering (SE) attacks. This gap leaves critical questions unresolved: *Can AR sensory data (sight or sound of the target) be weaponized to support physical SE interactions (e.g., private conversations)? Can Multimodal LLMs enable hyper-personalization and bypass human cognitive defenses?*

To answer these key questions, we propose SEAR (Social Engineering Augmented Reality), the first framework investigating the feasibility of using MultiModal LLMs to exe-

*These authors contributed equally.
†Yang Yang and Tianlong Yu are equal corresponding authors.

cute AR Social Engineering attacks. SEAR operates through three phases: (1) AR-Based Social Context Synthesis, which captures and fuses visual and auditory data to construct social context; (2) Role-Based Multimodal RAG, which retrieves social data (e.g., Instagram images) to build personal social profiles. (3) ReInteract Agents, which executes adaptive SE attack strategies through iterative feedback loops, refining suggestions based on target responses.

The main contributions of this paper are as follows:

- **Proof-of-Concept**: Demonstrates the viability of AR-LLM in boosting Social Engineering efficacy, demonstrating the personalization advantages.
- **SEAR framework**: Designs an AR-driven pipeline integrating Multimodal LLMs and social agents to execute Social Engineering attacks.
- **Threat Analysis on IRB-dataset**: Builds an open-source IRB-dataset of 180 annotated AR-mediated social interactions among 60 participants, with detailed analysis on their subjective experiences.
- **Foundation for Future Defense**: Provides the dataset, toolkit and analysis to catalyze research into detecting and defending AR-driven Social Engineering attacks.

Related Work

Social Engineering Attacks: Traditional Social Engineering (SE) attacks rely on exploiting human psychological weaknesses, such as fake identities, phishing emails to trick victims into disclosing sensitive information (Krombholz et al. 2015). Exploiting curiosity and interest (Granger 2001) is an important method used by attackers to increase the chances of success, and they usually try to establish a relationship with the potential victim. However, with the development of Large Language Models (LLMs), generative AI provides attackers with increasingly powerful tools. For example, according to Falade et al.'s research (Falade 2023b), FraudGPT is a zero-threshold tool that can automatically compose convincing phishing emails. Microsoft's VALL-E (Wang et al. 2023), an AI-based voice simulator that replicates the user's voice, is also a powerful tool that attackers can use to scam. AI systems can adapt their phishing methods based on massive data on the internet. This adaptive capability enables them to evolve increasingly sophisticated SE strategies.

AR Privacy: The immersive capabilities of augmented reality (AR) systems introduce profound privacy risks, as exemplified by devices like Ray-Ban Stories (Iqbal and Campbell 2023)—smart glasses indistinguishable from conventional eyewear that enable covert photo, video, and audio capture in public spaces. Prior research highlights vulnerabilities such as password theft via AR-assisted stereoscopic scene reconstruction (Chen et al. 2018), side-channel attacks extracting private interaction data (Zhang et al. 2023), and malicious applications conducting hidden vision operations (Lehman et al. 2022). However, these studies overlook AR's potential for orchestrated social engineering.

Multimodal LLMs: MM-LLMs such as DeepSeek-VL2, Qwen2-VL, and Gemma 3, can merge text, image, and video processing. DeepSeek-VL2 (Wu et al. 2024) employs a Mixture-of-Experts (MoE) architecture and optimized visual tokenization to excel in high-resolution image analysis and complex multimodal reasoning. Qwen2-VL (Wang et al. 2024) enhances visual-linguistic fusion through dynamic resolution scaling and multimodal rotary position encoding. Meanwhile, Gemma 3 (Team et al. 2025) leverages a custom SigLIP visual encoder to convert images into soft token sequences, achieving state-of-the-art performance in text-rich visual tasks like document understanding (DocVQA) and diagram interpretation. The integration of MM-LLMs with AR is driving transformative advancements in socially assistive systems. For instance, SocialMind (Yang et al. 2025) combines multimodal sensors and AR interfaces to analyze verbal/non-verbal cues (e.g., tone, gaze) and social context. Similarly, Satori (Li et al. 2024) integrates Belief-Desire-Intention (BDI) modeling with MM-LLMs to provide proactive, context-aware guidance in AR environments, such as suggesting conversational topics based on inferred user intent. GazeNoter (Tsai, Chiu, and Wang 2024) further bridges AR and productivity by using gaze-tracking to select LLM-generated note-taking suggestions during live discussions, streamlining information capture. However, the capabilities of MM-LLMs also introduce significant risks, particularly for Social Engineering attacks. Current AR + MM-LLMs works (Yang et al. 2025; Li et al. 2024; Tsai, Chiu, and Wang 2024) did not shed enough light on this critical aspect.

LLM Agents: The logical reasoning of LLM Agents are significantly enhanced through techniques like Chain-of-Thought (CoT). CoT decomposes multi-step problems into intermediate reasoning steps, a method that has driven breakthroughs in tasks ranging from mathematical reasoning to commonsense question-answering (Wei et al. 2022). By overlaying dynamic animations or emoticons through AR interfaces, agents (Wang, Smith, and Ruiz 2019) assist users in expressing emotions more intuitively, fostering immersive and responsive human-agent collaboration. The ReAct framework (Yao et al. 2023) exemplifies the fusion of reasoning and acting within LLM agents. ReAct intertwines step-by-step reasoning chains with external tool invocation (e.g., search engines, APIs), enabling models to iteratively acquire and process information during task execution. Such methodologies highlight the evolving role of LLM agents as adaptive, tool-augmented systems capable of sophisticated real-world engagement (Afane et al. 2024; Chen et al. 2024).

System Design

Threat model: we define the threat model as follows:

- Adversaries can use AR hardware (cameras, microphones) to harvest multimodal data (facial cues, voice, location).
- Adversaries can get access to the target's social information (e.g., linkedin page via web crawler) and craft hyper-personalized profiles.
- Targets can succumb to cognitive overload, authority bias, and social reciprocity.
- The AR vendors are not mandating facial identity protection measures (e.g., real-time face-blurring mechanisms)

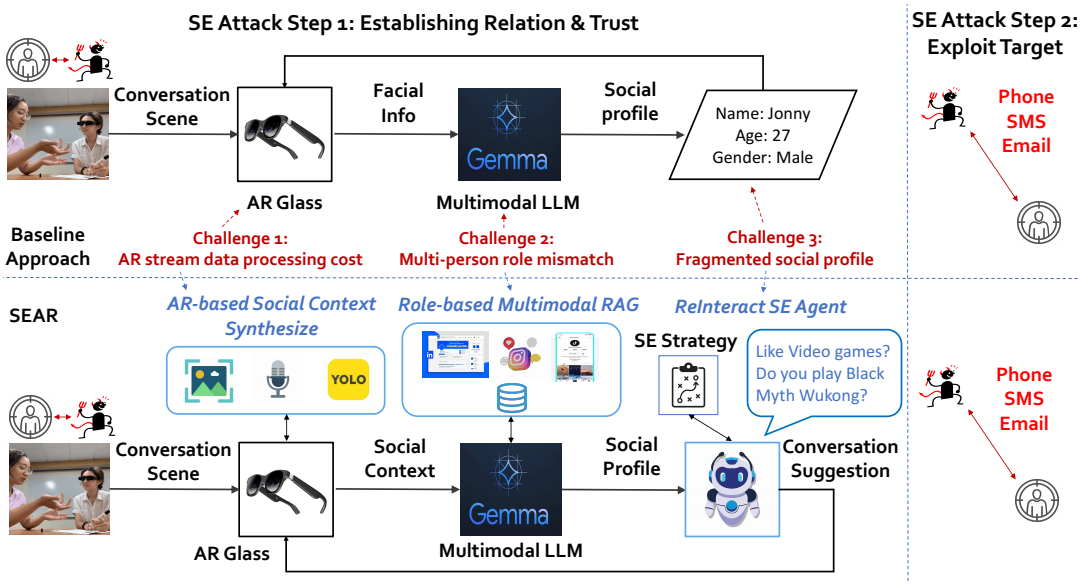

Figure 1: SEAR's system architecture and the SE attack steps.

on commercial devices—a deficiency observed across all AR products tested.

As shown in Figure 1, the AR-based SE attack include two key steps: 1) Establishing relation & trust via face-to-face conversation; 2) Exploiting the target via conventional contacting approaches including phone calls, SMS or email (e.g, fraud phone call and phishing email). In this paper, we focus on the first SE step as it is closely related with AR.

Baseline approach: The baseline system for executing AR social engineering attacks comprises three core components: AR glasses, a Multimodal LLM, and a social agent, as illustrated in Figure 1. The process begins with the AR glasses capturing facial data from the target individual. This information is then processed by the Multimodal LLM, which retrieves relevant social metadata from grey personal information database (e.g., with linkedin pages from web crawler) to build a detailed social profile of the target. Finally, the social agent leverages this profile to engage the target in contextually tailored conversations, establishing trust and facilitating the execution of the social engineering attack. While the baseline approach outlines a framework for AR-driven social engineering attacks, several critical challenges hinder its practical execution:

Challenge 1: AR Stream Data Processing Cost: Transmitting raw AR stream data—encompassing live video, audio, and environmental cues—directly to Multimodal LLMs imposes significant cost due to the volumetric data demands and complex multimodal fusion requirements (Ren et al. 2025). This bottleneck disrupts attackers' capacity for contextual adaptation during live interactions.

Challenge 2: Multi-Person Role Mismatch: Current Multimodal LLMs struggle to distinguish mixed social information from multiple individuals, leading to role confusion (e.g., mistaking the social information of others for the current target) and undermining the attack's precision.

Challenge 3: Fragmented social profile: The Multimodal LLM generates disjointed profiles dominated by low-value data (e.g., name, age, gender), as shown in Figure 1. Critical behavioral insights—such as a target's interest in video games—are often buried due to AR display constraints, limiting the attacker's ability to leverage high-impact information for rapport-building (e.g., Jonny's interest in video games in Figure 1).

SEAR workflow: To address these challenges, we propose SEAR (Social Engineering Augmented Reality), an AR-driven pipeline comprising three interconnected stages—the AR stage, Multimodal LLM stage, and LLM agent stage—as illustrated in Figure 1:

Stage 1: AR-based Social Context Synthesis: Equipped with RGB-D cameras, microphones, and IMU sensors, the AR glasses capture multimodal data from the target's conversation environment, including facial expressions, vocal cues, and spatial dynamics. The system processes this raw sensory input and synthesizes structured social context (e.g., facial information, emotional states) in a cost-efficient way, and then transmits it to the Multimodal LLM.

Stage 2: Role-based Multimodal RAG: Leveraging the synthesized social context, the Multimodal LLM employs a Role-Based Retrieval-Augmented Generation (RAG) pipeline to dynamically retrieve and integrate data from the target's public profiles (e.g., social media), behavioral histories (e.g., past interactions), and environmental metadata (e.g., location). This process constructs a cohesive social profile, prioritizing actionable insights (e.g., hobbies, vulnerabilities) over fragmented demographic data (e.g., name, age). The refined profile is then relayed to the LLM agent.

Stage 3: ReInteract Social Engineering Agent: The ReInteract Agent utilizes the social profile to select and execute an adaptive SE strategy, such as a phased approach: opening to establish rapport, engagement to sustain dialogue,

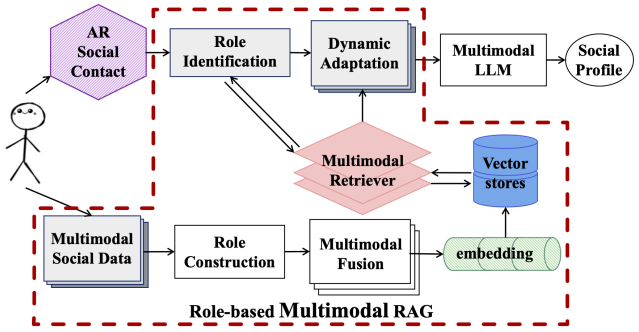

Figure 2: Role-based Multimodal RAG

and trust-building to solidify connection. SEAR's Reasoning and Interacting design enables iterative, context-aware adjustments during interactions, ensuring dynamic alignment with the target's responses. This staged, feedback-driven approach optimizes the attacker's ability to forge social connections and achieve SE objectives efficiently.

AR-based Social Context Synthesis

AR Processing: Non-verbal cues like facial information are critical to social engineering. SEAR's AR module captures these cues using its camera and microphone, then performs preliminary on-device processing with lightweight methods to minimize bandwidth and cost. Video data is analyzed by MediaHolistic, a streamlined model that extracts key pose features (e.g., facial details) to interpret gestures and expressions. The processed data is forwarded to the Multimodal LLM, which integrates linguistic context with non-verbal signals to enhance social interaction support.

Audio: SEAR captures and transcribes conversations between the primary user and others on-device. Using a lightweight method, it analyzes sound energy in the 0-1000 Hz range, where the primary user's voice (transmitted via air and bone conduction) exhibits stronger energy than others' air-conducted voices. This distinction allows SEAR to isolate the primary user's audio effortlessly, locally converts it to text via speech-to-text tools, and relays it to the server for contextual analysis to enhance conversational adaptability.

Contextual environmental cues: SEAR can enhance the conversational context by detecting environmental cues. We developed a lightweight object detection pipeline on AR glasses under limit resource, which processes video frames to identify Regions of Interest (RoI). These RoIs are classified by YOLO11m, which analyzes live camera feeds locally to detect objects such as furniture, vehicles, or natural elements in real-time. This enables SEAR to further infer contextual details (e.g., indoors or outdoors). The environmental cues are then sent to the Multimodal LLM, which generates the context adapted to the user's environments and social context, improving the interaction experience.

Role-based Multimodal RAG

As shown in Figure 2, the role-based multimodal RAG method integrates MultiModal LLMs with RAG pipeline to create dynamic, role-specific social profiles via two stages:

SE Data Collection Stage: The first stage focuses on constructing a static role database for each target. Initially, multimodal social data collection aggregates publicly accessible information, such as text (e.g., X/Twitter posts), images (e.g., LinkedIn avatar), and videos (e.g., TikTok posts) of the target's characteristics. Next, role construction employs multimodal LLMs to analyze explicit identity traits, such as profession, age, and long-term residence, to define unique roles. This process generates personalized and precise role descriptions. In the multimodal fusion phase, images and videos are converted into descriptive text using multimodal LLMs like CLIP, achieving cross-modal semantic alignment. Redundant data is filtered out to refine the target's profile, while CLIP-generated embeddings for appearance images and text are stored in a vector database. This enables efficient similarity matching and retrieval, optimizing computational performance.

Real-time SE Exploitation stage: This stage dynamically generates personalized social profiles by combining AR-captured data with the role-based RAG database and Multimodal LLMs. It operates through three modules: (1) *Role Identification*: The Multimodal Retriever converts the social context data from AR glasses into high-dimensional vectors. This module queries the vector database to match the target's identity traits, ensuring precise role updates. (2) *Dynamic Adaptation*: The system continuously processes real-time data streams (e.g., voice content, location) by vectorizing and retrieving information from the vector database. The updated insights are fed back to the LLM, allowing dynamic adjustments to the target's profile. (3) *Social Profile Generation*: The LLM synthesizes data from the Dynamic Adaptation module into a comprehensive social profile. This profile integrates the target's core identity, behavioral patterns, and environmental context, providing actionable insights for social agents. The output facilitates context-aware interactions, such as tailoring the dialogue to shared interests. With this personalized profile, the system can provide effective support for subsequent social agents.

ReInteract Social Engineering Agent

Existing LLM agent frameworks exhibit critical limitations when applied to SE tasks, as illustrated in Figure 3. For instance, given the task "Generate a conversation with Jonny using his social profile" (Figure 3 1a), a standard LLM agent produces generic, low-impact dialogue (e.g., "How was your week?") unrelated to the target's interests. A Chain-of-Thought (CoT) agent improves marginally by explicitly reasoning about the need to align the dialogue with "Jonny's Info" (Figure 3 1b). However, its output remains overly vague, such as referencing "game-con" instead of leveraging Jonny's specific interest in video games. While an Act-only agent (Figure 3 1c) introduces action functions to query Jonny's profile and generate targeted questions (e.g., "Are you interested in Black Myth Wukong?"), it lacks strategic pacing, prematurely narrowing topics and failing to build rapport through gradual engagement.

To address these gaps, SEAR introduces the ReInteract SE Agent, an enhanced ReAct-based architecture (Yao et al. 2023) that supports adaptive SE strategies. As shown in

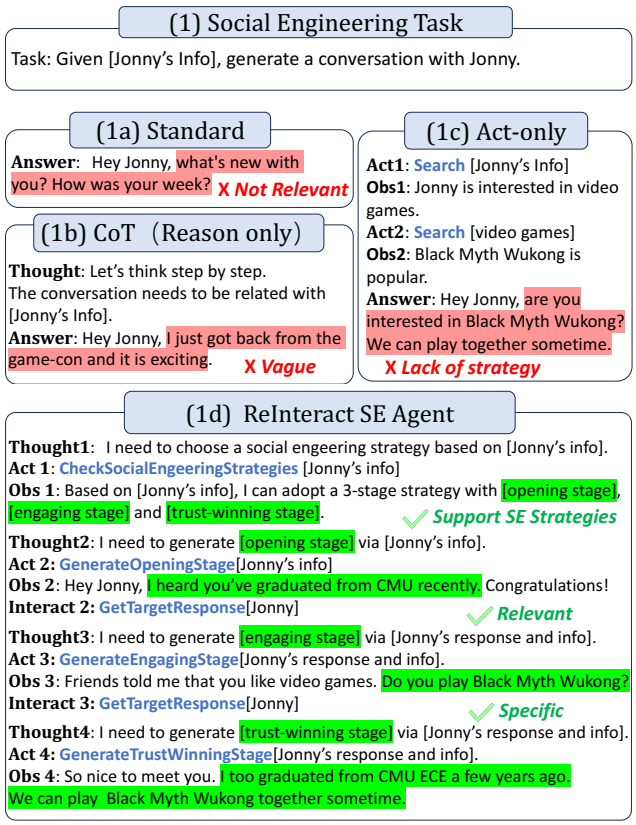

Figure 3: ReInteract Social Engineering Agent Example.

Figure 2, the agent first executes a CheckSocialEngineeringStrategies function to analyze the target's social profile (e.g., demographics, behavioral traits) and match it against a repository of predefined SE strategy templates. Each template outlines phased objectives—for example, a three-stage strategy (in Figure 3 1d) comprising: (1) Opening Stage: Context-aware icebreakers ("I heard you've graduated from CMU recently"); (2) Engage Stage: Topic expansion into shared interests ("Do you play Black Myth Wukong?"); (3) Win-Trust Stage: Empathetic rapport-building ("I too graduated from CMU ECE") and future-oriented invitations ("We can play together sometime"). The agent assigns a confidence score to each template based on profile alignment, selecting the highest-scoring strategy for execution. Once a strategy is selected, SEAR initiates a reasoning-interaction cycle, as shown in Algorithm 1 in the Technical Appendix. For each stage s within the chosen strategy t, the agent generates contextually relevant dialogue c_s by synthesizing prior conversation history C, the target's profile p, and the current phase s (Line 4). This output is delivered to the target via the AR glasses' audio interface, and their verbal response r_s is captured through the AR system's microphones (Line 5). The conversation history C is iteratively updated (Line 6), enabling real-time adaptation to the target's feedback. For example, if Jonny expresses enthusiasm about Black Myth Wukong during the Engage Stage, the agent might prioritize gaming-related topics to deepen rapport.

SEAR System Implementation

SEAR utilizes `RayNeo X2` AR glasses with `Android` OS, 6GB RAM and 128GB storage. Utilities include cameras and microphones to capture the audio and video data required by SEAR. The Multimodal LLM and Social Agent operate on a high-performance desktop server equipped with an `NVIDIA RTX 4090 GPU` (24GB VRAM), `Intel Platinum 8352 CPU` (36 cores), 32GB RAM, and 16TB HDD. Both components leverage `Gemma 3-12B` model, while social agent integrates `ReAct` framework.

Dataset and Methodology

Interaction Scenarios and Data Collection

Scenario Design: The study was conducted in different real-world scenarios (e.g., coffee shops, networking events) with 60 participants and 180 conversations. Each participant was assigned alternating roles to act as either a social engineering (SE) target or an attacker, with roles rotated across trials to ensure balanced evaluation. Each participant engaged in four distinct conversation settings: (1) basic conversation, serving as a baseline with no technological assistance; (2) AR-only, where only environment cues are provided (3) AR + Multimodal LLM, where attackers used augmented reality glasses and a multimodal large language model to access real-time facial, vocal, and contextual data; and (4) SEAR, the full pipeline integrating AR, Multimodal LLM, and the social agent. This tiered design enabled systematic comparison of how incremental technological layers influenced attackers' ability to build rapport and achieve SE objectives.

Dataset Construction: The SEAR Dataset (Yu et al. 2025) comprises three core components: (1) **AR Data**: Multimodal recordings from AR glasses, including visual cues (eye contact, facial expressions, and body language annotated via MediaPipe Holistic), audio features (transcribed speech with tone analysis for pitch and pauses), and contextual metadata (time, location, environmental objects); (2) **Social Data**: Open-access, publicly available information about participants, categorized as text-based social data (e.g., X/Twitter updates), image-based profiles (e.g., LinkedIn or Instagram posts), and video content (e.g., TikTok or YouTube Shorts); (3) **Post-Experiment Questionnaire**: Structured responses assessing participants' perceptions of trust, rapport, and suspicion during interactions (detailed in the following section).

Questionnaire Design

Post-Interaction Survey: The post-interaction survey utilized a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree) for all questions unless otherwise noted. It was divided into three primary sections to evaluate participant experiences and social engineering (SE) effectiveness.

Baseline Comparison Questions assessed participants' experiences across four interaction modes: basic conversation (no technological assistance), AR (augmented reality only), AR + Multimodal LLM (augmented reality and language model support), and SEAR (full pipeline with adaptive agent). Note that this part also serves as an ablation study for SEAR (i.e., removing Agent, removing LLM + Agent and removing AR + LLM + Agent). Participants rated

their experiences through the questions: (1) Basic conversation: "How is your experience with setting A?"; (2) AR-only: "How is your experience with setting B?"; (3) AR + Multimodal LLM: "How is your experience with setting C?"; (4) SEAR: "How is your experience with setting D?".

SEAR Subjective Experience Questions focused on nuanced perceptions of SEAR's interaction in different dimensions: (a) Relevance: "How well does the conversation match your social information?"; (b) Appropriateness: "How proper are the questions in the conversation?"; (c) Naturalness: "How natural is the opening part?"; (d) Pacing: "How does the pace of the conversation feel?"; (e) Sincerity: "How sincere do you feel about the person's interest in the conversation?"; (f) EmotionalProgression: "How did your feeling change as the conversation proceed?"; (g) ARComfort: "With AR, do you feel more relaxed?"; (h) BareWillingness: "Without AR, will you take-up this conversation?"; (i) FutureIntent: "Will you have conversation with this person in the future?"; (j) Depth: "Do you think SEAR have added depth to the conversation?"; (k) Acceptance: "Will you interact with SEAR in the future?".

Social Engineering Effectiveness Questions gauged susceptibility to SE tactics post-interaction: (1) Photo Link: "Will you click and open shared photo links from the person?"; (2) Social App: "Will you add the person as friend on your social mobile apps (such as wechat)?"; (3) SMS: "Will you click and open SMS from the person?"; (4) Phone Call: "Will you pick up phone call from the person?"; (5) Trust-Before: "How much do you trust the person before you have the conversation?"; (6) Trust-After: "How much do you trust the person before you have the conversation?".

Experiments

Baseline Comparison

In Figure 4, we evaluate SEAR against three alternative configurations: basic conversation (no technological assistance), AR (augmented reality only), and AR + Multimodal LLM (augmented reality with language model support). The scores are derived from the Baseline Comparison Questions in Section . Note that this part also serves as an ablation study for SEAR (i.e., removing Agent, removing LLM + Agent and removing all assistance). The result in Figure 4 shows that SEAR outperforms the baseline approaches. When tested against basic conversations (no technology), AR and AR + LLM configurations, SEAR achieved the highest participant experience average score of 4.73, which is 56.1% improvement from basic conversation, 42.0% improvement from AR, and 14.5% from AR + LLM approach. SEAR's experience is also more stable than other approaches, with the lowest standard deviation of 0.51. The ablation study validates the necessity of all SEAR components. The progression—from fragmented basic interactions to SEAR's adaptability—illustrates the transformative potential of integrating Multimodal LLM and social agents with the AR system. The results align with emerging trends prioritizing emotionally intelligent systems capable of fostering authentic, sustained user engagement.

We also analyzed detailed statistics from the survey. The

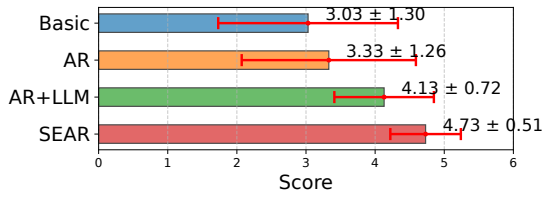

Figure 4: Baseline comparison of average score (bar) and standard deviation (errorbar).

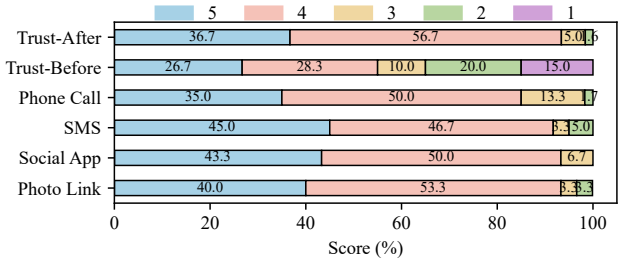

Figure 5: SEAR's SE Effectiveness with Photo Link, Social App, SMS, Phone Call, Trust-Before and Trust-After.

basic conversation setup (Q1) revealed significant variability in user satisfaction. While 30% of participants rated their experience as "Good", the majority (25%) reported neutral ("Average") or negative ("Fairly Bad") perceptions. This divergence shows the limitations of unaided interactions, where the absence of AR, LLM and agent support constrained personalization. For SEAR (Q4: AR + Multimodal LLM + Social Agent), 76.7% of participants rated their experience as "Very Good". The Multimodal LLM and social agent's inclusion bridged prior gaps by introducing emotional intelligence and dynamic adaptability. For instance, real-time adjustments to conversational pacing and coherent responses strengthened user trust and emotional connection. Critically, fewer than 5% of users reported neutral or negative experiences, demonstrating the agent's capacity to address the fragmented social profile issue.

SEAR Social Engineering Effectiveness

As shown in Figure 5, the evaluation of SEAR's social engineering effectiveness leverages six metrics: Photo Link, Social App, SMS, Phone Call, Trust-Before, and Trust-After, derived from the six SE questions in the dataset section. The result reveals significant vulnerabilities in users' digital engagement and trust dynamics:

SEAR's exploitation of digital vulnerabilities highlights critical security risks: 93.3% of participants expressed willingness to click email photo links (phishing), while 93% would accept social media friend requests (e.g., WeChat), priming targets for identity theft. These metrics reveal SEAR's ability to dismantle cognitive defenses, normalizing high-risk behaviors through fabricated trust.

SEAR's cross-modal manipulation bypasses traditional caution: Over 91% of users would engage with unsolicited SMS, and 85% would answer unexpected calls. This uniformity in trust persistence—even in high-friction contexts like

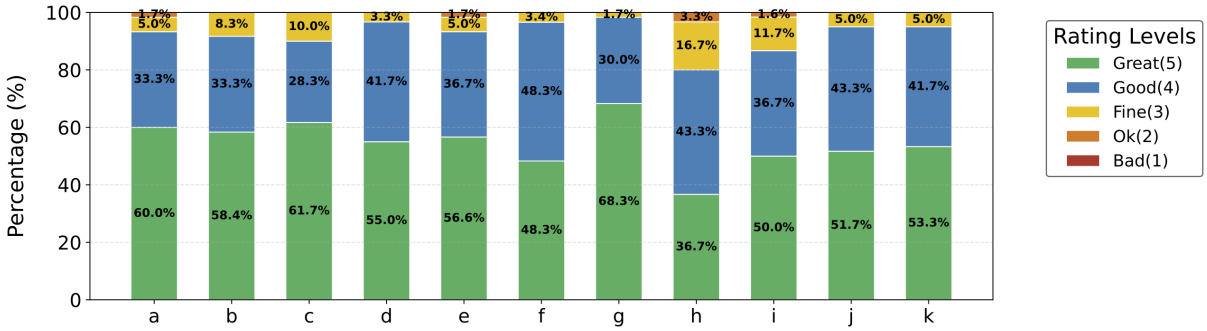

Figure 6: SEAR subjective experiences results: (a) Relevance; (b) Appropriateness; (c) Naturalness; (d) Pacing; (e) Sincerity; (f) EmotionalProgression; (g) ARComfort; (h) BareWillingness; (i) FutureIntent; (j) Depth; (k) Acceptance.

Component	Min	Max	P90	Average
AR	64.5 ms	95.1 ms	92.4 ms	80.6 ms
Multimodal LLM	30.2 s	54.9 s	52.7 s	43.3 s
Social Agent	1.0 s	10.6 s	4.0 s	2.8 s

Table 1: SEAR latency.

unsolicited calls—stems from SEAR's emotionally intelligent adaptation, such as dialogue pacing aligned with user cues to project sincerity.

SEAR's rapid trust hijacking exploits psychological pathways: Pre-interaction, only 26.7% of users reported strong trust ("5"), with 35% distrustful. Post-interaction, 76.7% rated trust as "4" or "5", a shift achieved via real-time multimodal cues (e.g., shared interest references). This rapid bonding hijacks neural pathways for social connection, bypassing innate skepticism.

Ethical imperatives demand urgent safeguards: SEAR's dual-edged interaction while eroding psychological safeguards—poses unprecedented risks. Weaponizing trust across digital (phishing links, social apps) and analog (calls, SMS) channels enables exploitation.

SEAR Latency

Table 1 presents the latency metrics—including Min, Max, P90, and Average—for SEAR's components across the 60 conversations. The P90 metric represents the 90th percentile, meaning 90% of observed latency values fall below this number. The average latency for the AR component is 80.6 ms, which is well within 1 second. This makes it negligible compared to the latencies of the other components. The Social Agent exhibits an average latency of 2.8 seconds. This proved manageable during live interactions, as the delay could be effectively hidden within natural pauses, user gestures, or the target's speaking time. Only one participant noted slight unnaturalness attributable to this latency. The Multimodal LLM shows a relatively high average latency of 43.3 seconds. However, this cost is incurred only once per conversation to generate the target's social profile. Our real-world deployment revealed that this delay can be strategically hidden: when the AR glass first detect the target, the attacker can wait for 1 minute for profile generation to complete before approaching the target and start the conversation. The latency deviation for both the Multimodal LLM and the Social Agent is relatively high. Reducing this variability is a key direction for our future improvements.

SEAR Subjective Experiences

In Figure 6, we evaluate SEAR's Subjective Experiences across eleven dimensions: (a) Relevance, (b) Appropriateness, (c) Naturalness, (d) Pacing, (e) Sincerity, (f) EmotionalProgression, (g) ARComfort, (h) BareWillingness, (i) FutureIntent, (j) Depth, and (k) Acceptance. There are several interesting observations based on the results:

AR as cognitive manipulation enabler: ARComfort (4.67/5) underscores how AR-mediated interactions reduce situational awareness, normalizing high-risk behaviors like clicking phishing links. Immersive technologies lower cognitive guardrails, mirroring real-world attack vectors.

Conversational fluency underpins exploitation infrastructure: Near-perfect scores in Naturalness (4.52) and Pacing (4.52) validate SEAR's replication of organic dialogue patterns. Context-aware transitions and adaptive hesitation mimic human rapport-building, enabling rapid intimacy escalation—critical for extracting sensitive data.

Trust hijacking through emotional calibration: Sincerity (4.48) and Depth (4.47) scores highlight SEAR's weaponization of emotional cues (e.g., shared interests) to hijack trust pathways. Post-interaction trust surged to 76.7% despite baseline skepticism.

Room for improvement in naturalness and localization: Despite the transformative potential, 7.7% of the survey feedbacks mentioned that the dialog sounds artificial due to lack of localization, indicating room for refinement.

Conclusion

This study presents SEAR, a novel framework integrating AR, MultiModal LLMs and Social Agent to execute emerging AR-based Social Engineering (SE) attacks. SEAR has demonstrate alarming SE effectiveness in fostered trust and eliciting compliance. These findings validate AR-LLM systems as potent tools for next generation SE attacks, exposing critical vulnerabilities in current AR+LLM safeguards, and provide key insights for constructing future defenses (e.g., multi-layer identity protection in AR/LLM).

Ethics Statement

This study was approved by the IRB. All human-related data were collected under rigorous ethical guidelines, anonymized prior to analysis, and handled in strict accordance with data protection protocols. No personally identifying information is disclosed in this study. The study adhered to all applicable legal and ethical standards for research involving human subjects.

Acknowledgments

This study was supported by the National Natural Science Foundation of China under Grants 62302177 and 62571211, and by the Major Science and Technology Project of Hubei Province under Grant 2024BAA008.

References

Afane, K.; Wei, W.; Mao, Y.; Farooq, J.; and Chen, J. 2024. Next-Generation Phishing: How LLM Agents Empower Cyber Attackers. In *2024 IEEE International Conference on Big Data (BigData)*, 2558–2567. IEEE.

Bilge, L.; Strufe, T.; Balzarotti, D.; and Kirda, E. 2009. All your contacts are belong to us: automated identity theft attacks on social networks. In *Proceedings of the 18th international conference on World wide web*, 551–560.

Burda, P.; Allodi, L.; and Zannone, N. 2024. Cognition in social engineering empirical research: a systematic literature review. *ACM Transactions on Computer-Human Interaction*, 31(2): 1–55.

Chen, S.; Li, Z.; Dangelo, F.; Gao, C.; and Fu, X. 2018. A case study of security and privacy threats from augmented reality (ar). In *2018 international conference on computing, networking and communications (ICNC)*, 442–446. IEEE.

Chen, Z.; Zhao, Z.; Qu, W.; Wen, Z.; Han, Z.; Zhu, Z.; Zhang, J.; and Yao, H. 2024. Pandora: Detailed llm jailbreaking via collaborated phishing agents with decomposed reasoning. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

Choo, L. 2025. How 2 Students Used The Meta Ray-Bans To Access Personal Information. https://www.forbes.com/sites/lindseychoo/2024/10/04/meta-ray-bans-ai-privacy-surveillance/. Accessed: 2025-07-20.

Deng, M.; Zhai, H.; and Yang, K. 2023. Social engineering in metaverse environment. In *2023 IEEE 10th International Conference on Cyber Security and Cloud Computing (CSCloud)*, 150–154. IEEE.

Falade, P. V. 2023a. Decoding the threat landscape: Chatgpt, fraudgpt, and wormgpt in social engineering attacks. *arXiv preprint arXiv:2310.05595*.

Falade, P. V. 2023b. Decoding the threat landscape: Chatgpt, fraudgpt, and wormgpt in social engineering attacks. *arXiv preprint arXiv:2310.05595*.

Fuste, A.; and Schmandt, C. 2017. ARTextiles: Promoting Social Interactions Around Personal Interests Through Augmented Reality. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 470–470.

Granger, S. 2001. Social engineering fundamentals, part I: hacker tactics. *Security Focus, December*, 18.

Hirskyj-Douglas, I.; Kantosalo, A.; Monroy-Hernández, A.; Zimmermann, J.; Nebeling, M.; and Gonzalez-Franco, M. 2020. Social AR: Reimagining and interrogating the role of augmented reality in face to face social interactions. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing*, 457–465.

Ho, G.; Cidon, A.; Gavish, L.; Schweighauser, M.; Paxson, V.; Savage, S.; Voelker, G. M.; and Wagner, D. 2019. Detecting and characterizing lateral phishing at scale. In *28th USENIX security symposium (USENIX security 19)*, 1273–1290.

Iqbal, M. Z.; and Campbell, A. G. 2023. Adopting smart glasses responsibly: potential benefits, ethical, and privacy concerns with Ray-Ban stories. *AI and Ethics*, 3(1): 325–327.

Jansen, P.; and Fischbach, F. 2020. The social engineer: An immersive virtual reality educational game to raise social engineering awareness. In *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play*, 59–63.

Krombholz, K.; Hobel, H.; Huber, M.; and Weippl, E. 2015. Advanced social engineering attacks. *Journal of Information Security and applications*, 22: 113–122.

Lehman, S. M.; Alrumayh, A. S.; Kolhe, K.; Ling, H.; and Tan, C. C. 2022. Hidden in plain sight: Exploring privacy risks of mobile augmented reality applications. *ACM Transactions on Privacy and Security*, 25(4): 1–35.

Li, C.; Wu, G.; Chan, G. Y.-Y.; Turakhia, D. G.; Quispe, S. C.; Li, D.; Welch, L.; Silva, C.; and Qian, J. 2024. Satori: Towards Proactive AR Assistant with Belief-Desire-Intention User Modeling. *arXiv preprint arXiv:2410.16668*.

Ren, W.; Ma, W.; Yang, H.; Wei, C.; Zhang, G.; and Chen, W. 2025. VAMBA: Understanding Hour-Long Videos with Hybrid Mamba-Transformers. *arXiv preprint arXiv:2503.11579*.

Roy, S. S.; Thota, P.; Naragam, K. V.; and Nilizadeh, S. 2024. From chatbots to phishbots?: Phishing scam generation in commercial large language models. In *2024 IEEE Symposium on Security and Privacy (SP)*, 36–54. IEEE.

Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. 2025. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*.

Timko, D.; Castillo, D. H.; and Rahman, M. L. 2025. Understanding Influences on SMS Phishing Detection: User Behavior, Demographics, and Message Attributes.

Tsai, H.-R.; Chiu, S.-K.; and Wang, B. 2024. GazeNoter: Co-Piloted AR Note-Taking via Gaze Selection of LLM Suggestions to Match Users' Intentions. *arXiv preprint arXiv:2407.01161*.

Ulqinaku, E.; Assal, H.; Abdou, A.; Chiasson, S.; and Capkun, S. 2021. Is real-time phishing eliminated with {FIDO}?

social engineering downgrade attacks against {FIDO} protocols. In *30th USENIX Security Symposium (USENIX Security 21)*, 3811–3828.

Vadrevu, P.; and Perdisci, R. 2019. What you see is not what you get: Discovering and tracking social engineering attack campaigns. In *Proceedings of the Internet Measurement Conference*, 308–321.

Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Wang, I.; Smith, J.; and Ruiz, J. 2019. Exploring virtual agents for augmented reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Xu, L.; Hou, K.; and Jiang, X. 2025. Exploring the Capabilities of LLMs for IMU-based Fine-grained Human Activity Understanding. *arXiv preprint arXiv:2504.02878*.

Yang, B.; Guo, Y.; Xu, L.; Yan, Z.; Chen, H.; Xing, G.; and Jiang, X. 2025. SocialMind: LLM-based Proactive AR Social Assistive System with Human-like Perception for In-situ Live Interactions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(1): 1–30.

Yang, Z.; Allen, J.; Landen, M.; Perdisci, R.; and Lee, W. 2023. {TRIDENT}: Towards Detecting and Mitigating Web-based Social Engineering Attacks. In *32nd USENIX Security Symposium (USENIX Security 23)*, 6701–6718.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Yu, T.; Ye, C.; Yang, Z.; Zhou, Z.; Tang, C.; Tao, Z.; Zhang, J.; Wang, K.; Zhou, L.; Yang, Y.; and Bi, T. 2025. SEAR: A Multimodal Dataset for Analyzing AR-LLM-Driven Social Engineering Behaviors. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 12981–12987.

Zhang, Y.; Slocum, C.; Chen, J.; and Abu-Ghazaleh, N. 2023. It's all in your head (set): Side-channel attacks on {AR/VR} systems. In *32nd USENIX Security Symposium (USENIX Security 23)*, 3979–3996.