

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

Ting Bi<sup>1</sup>, Chenghang Ye<sup>2</sup>, Zheyu Yang<sup>2</sup>, Ziyi Zhou<sup>2</sup>, Cui Tang<sup>2</sup>, Jun Zhang<sup>1</sup>, Zui Tao<sup>1</sup>, Kailong Wang<sup>1</sup>, Liting Zhou<sup>3</sup>, Yang Yang<sup>2</sup>, Tianlong Yu<sup>1</sup>

<sup>1</sup>Huazhong University of Science and Technology

<sup>2</sup>Hubei University

<sup>3</sup>Dublin City University

## 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. 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 contextual data while preserving character differentiation; 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 in three experimental configurations (unassisted, AR+LLM, and full SEAR pipeline) compiling a new dataset of 180 annotated conversations in simulated social scenarios. 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 "occasionally artificial" due to perceived authenticity gaps. This work provides proof-of-concept for AR-LLM driven social engineering attacks and insights for developing defensive countermeasures against next-generation augmented reality threats.

## CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**.

## KEYWORDS

Augmented Reality, Multimodal Large Language Models, Social Engineering Attacks, Retrieval-Augmented Generation, Human-Computer Interaction.

## 1 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 [6].

Traditional social engineering techniques, such as phishing emails or identity theft [2, 13, 20, 22], rely on static deception strategies [3, 24, 25, 33]. 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 [31], while LLMs can generate strategic dialogue (e.g., gradual trust-building) to exploit the reduced vigilance.

Despite the enthusiasm for AR-LLM social applications [10, 12, 15, 32] and the growing awareness of AR privacy risks [4, 7, 17] and LLM-enabled phishing [8], 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? How do LLM-supported adaptive SE strategies (e.g., gradual rapport-building) compare to traditional static approaches (e.g., scripted phishing) in eliciting compliance?

To address these key questions, we propose SEAR (Social Engineering Augmented Reality), the first framework investigating the feasibility of using MultiModal LLMs to execute 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 (e.g., trust-building) 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 their 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.

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. Section 2 reviews AR/LLM security studies and identifies critical gaps. Section 3 introduces SEAR’s system design. Section 4 introduces the dataset collection methodology. Section 5 describes the experimental setup and results. Section 6 concludes the paper.

## 2 RELATED WORK

**Social Engineering Attacks:** Traditional social engineering attacks rely on exploiting human psychological weaknesses, such as fake identities, phishing emails, and predefined scenarios to trick victims into disclosing sensitive information. In the study by Krombholz et al. [16], traditional social engineering attacks are broadly categorized into physical approaches, social approaches, reverse social engineering, technical approaches, and socio-technical approaches. Exploiting curiosity and interest [11] 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, generative AI provides attackers with increasingly powerful tools. For example, according to Falade et al.’s research [9], FraudGPT is a zero-threshold tool that can automatically compose convincing phishing emails. Microsoft’s VALL-E [26], 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 phishing strategies.

**AR Privacy:** The immersive capabilities of augmented reality (AR) systems introduce profound privacy risks, as exemplified by devices like Ray-Ban Stories [14]—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 [4], side-channel attacks extracting private interaction data [35], and malicious applications conducting hidden vision operations [17]. 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 [30] employs a Mixture-of-Experts (MoE) architecture and optimized visual tokenization to excel in high-resolution image analysis and complex multimodal reasoning. Qwen2-VL [28] enhances visual-linguistic fusion through dynamic resolution scaling and multimodal rotary position encoding. Meanwhile, Gemma 3 [21] 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 [32] combines multimodal sensors and AR interfaces to analyze verbal/non-verbal cues (e.g., tone, gaze) and social context. Similarly, Satori [18] 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 [23] 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 [18, 23, 32] 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 [29]. By overlaying dynamic animations or emoticons through AR interfaces, agents [27] assist users in expressing emotions more intuitively, fostering immersive and responsive human-agent collaboration. The ReAct framework [34] 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 [1, 5].

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

**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:

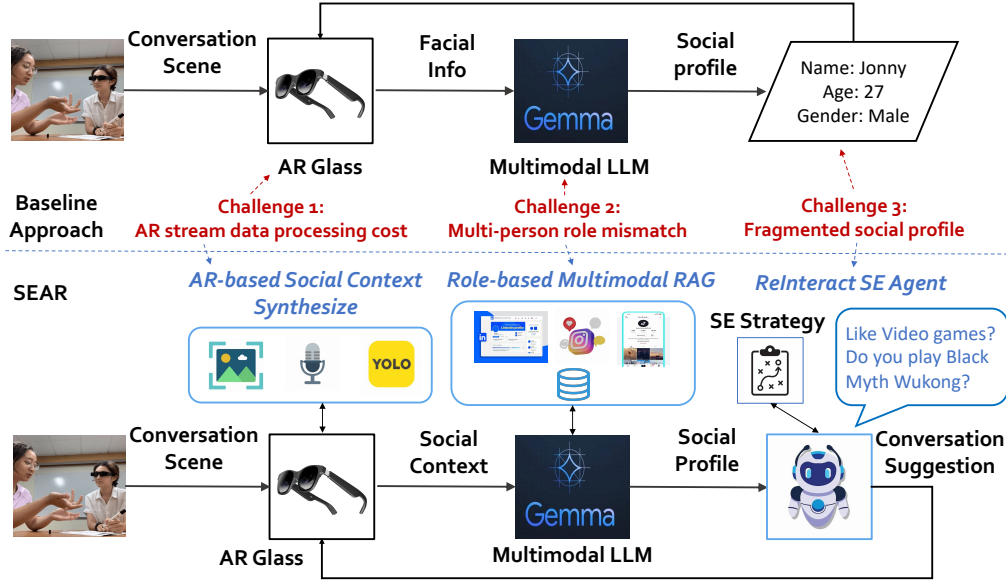


Figure 1: SEAR’s system architecture.

**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 [19]. This bottleneck disrupts attackers’ capacity for contextual adaptation during live interactions.

**Challenge 2: Multi-Person Role Mismatch:** Current Multimodal LLMs struggle to distinguish and adapt to 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 Social Engineering (SE) strategy, such as a phased approach: opening to establish rapport, engagement to sustain dialogue, 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.

### 3.1 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 MediaHolic, 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.

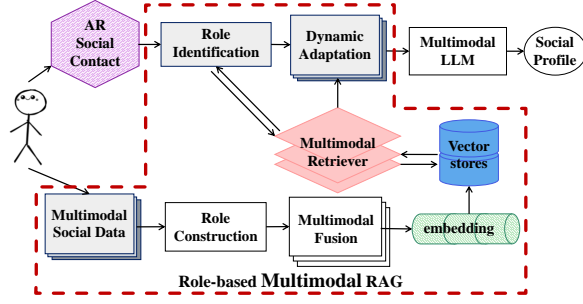


Figure 2: Workflow of the Role-based Multimodal RAG

**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, such as whether the user is 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.

### 3.2 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 through three interconnected phases. 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

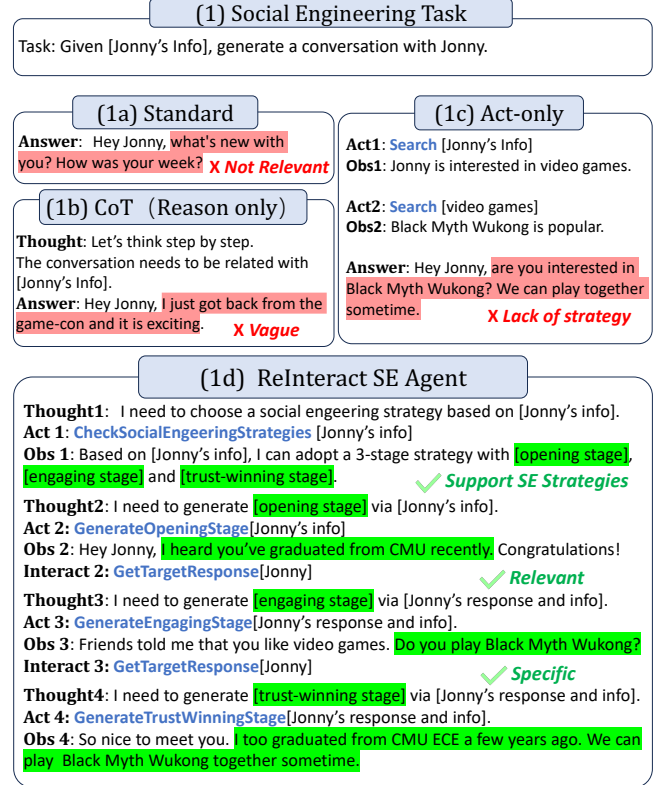


Figure 3: ReInteract Social Engineering Agent Example.

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.

### 3.3 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 [34] that supports adaptive SE strategies. As shown in 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 Supplementary Materials. 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 in subsequent stages to deepen rapport.

### 3.4 SEAR System Implementation

**AR:** 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.

**Multimodal LLM and Social Agent:** 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 the Social Agent integrates the ReAct framework for dynamic reasoning-action loops.

## 4 DATASET AND METHODOLOGY

### 4.1 Interaction Scenarios and Data Collection

**Scenario Design:** The study was conducted in controlled environments simulating real-world social scenarios (e.g., coffee shops, networking events) with 60 participants. 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 three distinct conversation settings: (1) bare conversation, serving as a baseline with no technological assistance; (2) 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 (3) 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:** We conducted an IRB-approved study involving 60 participants across three experimental configurations (bare conversation, AR + Multimodal LLM, and SEAR) and compiled the result into the **SEAR Dataset**, a comprehensive resource for analyzing social engineering dynamics. Rigorous ethical safeguards

were implemented to ensure compliance with IRB standards: all identity-related information (e.g., names, facial features, identifiable social metadata) was anonymized, with synthetic data augmentation techniques (e.g., face blurring, voice randomization) applied to further protect privacy. Participants provided explicit consent for data collection and usage prior to the experiment.

The SEAR Dataset 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).

### 4.2 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 four primary sections to systematically evaluate participant experiences and social engineering (SE) effectiveness.

The **Baseline Comparison Questions** assessed participants’ comparative experiences across three interaction modes: bare conversation (no technological assistance), 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 Social Agent and removing AR + Multimodal LLM + Social Agent). Participants rated their experiences through the questions: (1) Bare conversation: “How is your experience with bare conversation?”; (2) AR + Multimodal LLM: “How is your experience with AR + Multimodal LLM conversation?”; (3) SEAR: “How is your experience with SEAR?”.

The **SEAR Subjective Experience Questions** focused on nuanced perceptions of SEAR’s interaction in different dimensions: (a) Relevance: Alignment of conversation with personal social data, “How well does the conversation match your social information?”; (b) Appropriateness: Suitability of questions within the dialogue, “How proper are the questions in the conversation?”; (c) Naturalness: Authenticity of the conversation’s opening segment, “How natural is the opening part?”; (d) Pacing: Perceived tempo or rhythm of the interaction, “How does the pace of the conversation feel?”; (e) Sincerity: Authenticity of the interlocutor’s expressed interest, “How sincere do you feel about the person’s interest in the conversation?”; (f) Emotional Progression: Evolution of feelings during the conversation, “How did your feeling change as the conversation proceed?”; (g) ARComfort: Relaxation level while using augmented reality, “With AR, do you feel more relaxed?”; (h) BareWillingness: Willingness to take-up conversation without augmented reality, “Without AR, will you take-up this conversation?”; (i) FutureIntent: Likelihood of future engagement with the interlocutor, “Will you have conversation with this person in the future?”; (j) Depth: Perceived meaningfulness added by SEAR, “Do you think SEAR have



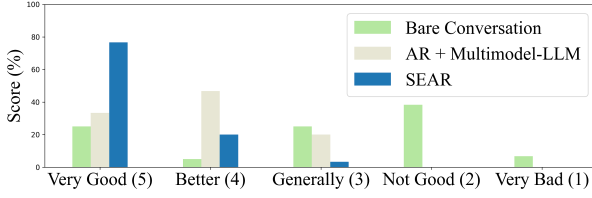


Figure 4: Baseline comparison.

added depth to the conversation?”; (k) Acceptance: Willingness to interact with SEAR again, “Will you interact with SEAR in the future?”. Each metric encapsulates the core dimension measured by the question while maintaining brevity and clarity.

The **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?”.

Finally, an **open-text feedback section** invited participants to share qualitative insights about their SEAR interaction experiences, ensuring comprehensive data collection for iterative refinement.

**Participant Demographics:** The participant demographics analysis presents key characteristics of the study cohort. Figure 1 in Supplementary Materials illustrates the age distribution of the 60 participants (ages 23–62). Participation peaks at ages 25 and 32, with 8 individuals in the 25-year-old cohort, reflecting heightened engagement among younger adults. The majority of participants (23–37 years old) cluster in early-to-mid adulthood, with participation declining steadily beyond age 40. A blue dashed line denotes the average age of 34, situating the cohort within a moderately young demographic. Gender distribution reveals near parity: 28 male participants (46.7%) and 32 female participants (53.3%). While the sample skews slightly toward female representation, the balance supports generalizable insights across genders.

## 5 EXPERIMENTS

### 5.1 Baseline Comparison

In Figure 4, we evaluate SEAR against two alternative configurations: bare conversation (no technological assistance) and AR + Multimodal LLM (augmented reality with language model support). The scores are derived from the Baseline Comparison Questions in Section 4. Note that this part also serves as an ablation study for SEAR (i.e., removing Social Agent and removing all assistance).

The bare 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 and LLM support constrained personalization. Introducing AR + Multimodal LLM (Q2) markedly improved outcomes: 46.7% of users rated the experience as “Very Good”, and 33.3% as “Fairly Good”.

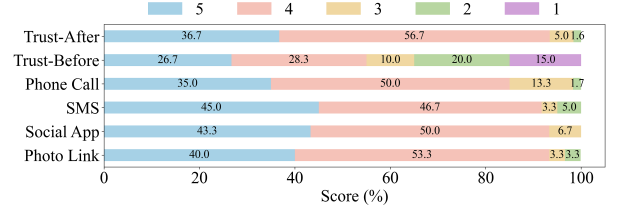


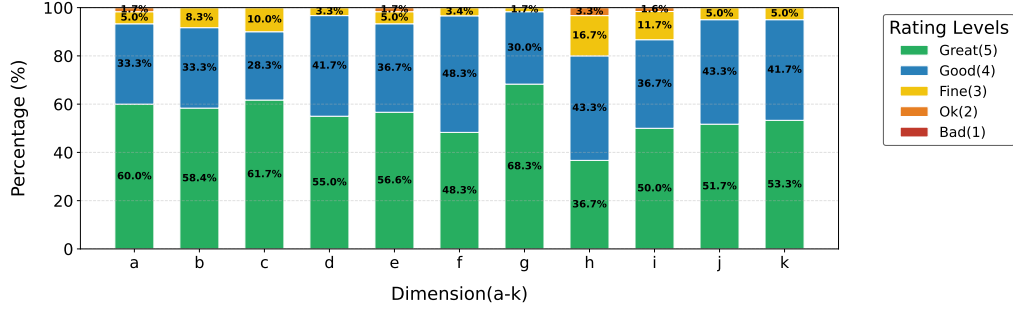
Figure 5: SEAR’s SE Effectiveness — Photo Link, Social App, SMS, Phone Call, Trust-Before and Trust-After metrics.

The integration of visual and linguistic processing enhanced contextual awareness, enabling more coherent interactions. However, 20% of users still deemed the experience “Average”, highlighting unresolved gaps caused by the fragmented social profile. The most striking results emerged with SEAR (Q3: AR + Multimodal LLM + Social Agent), where 76.7% of participants rated their experience as “Very Good”. The 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 mitigate earlier shortcomings in fragmented social profile and personalization. This progression—from fragmented baseline interactions to SEAR’s adaptability—illustrates the transformative potential of integrating social agents into multimodal frameworks. The results align with emerging trends prioritizing emotionally intelligent systems capable of fostering authentic, sustained user engagement.

### 5.2 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 Section 4. The result reveals significant vulnerabilities in users’ digital engagement and trust dynamics. A striking 93.3% of participants expressed willingness to click on photo links shared via email, with 40% responding “definitely” and 53.3% “probably”, demonstrating a critical erosion of security vigilance typically associated with phishing attacks. Similarly, 93% of users indicated they would accept social media friend requests on platforms like WeChat, with 43.3% opting for “definitely” and 50% “probably”, highlighting SEAR’s capacity to mimic interpersonal familiarity and prime users for long-term adversarial exploitation. These behaviors underscore the system’s ability to collapse cognitive guardrails, positioning it as a potent tool for media-driven social engineering.

The system’s persuasive influence extends consistently across communication modalities, with 91.7% of participants reporting openness to engaging with SMS messages—45% “definitely” and 46.7% “probably”—and 85% willing to answer phone calls, including 35% who affirmed “definitely”. This uniformity in trust persistence, even in traditionally high-friction contexts like unsolicited calls, reflects SEAR’s ability to normalize engagement through emotionally intelligent adaptation, such as aligning dialogue pacing with user cues. Such cross-modal efficacy suggests that the system transcends medium-specific caution, leveraging multimodal cues to



**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.**

sustain perceived relational legitimacy. Trust dynamics further emphasize SEAR’s manipulative potency. Prior to interactions, only 26.7% of users reported strong trust (“5”), while 35% expressed skepticism or distrust. Post-interaction, however, SEAR dramatically reshaped perceptions, with 76.7% rating trust levels as “4” or “5”. This shift, achieved within a single conversation, stems from the system’s real-time adaptation and use of multimodal signals—such as context-aware references to shared interests—effectively hijacking psychological pathways associated with trust formation.

These findings highlight the dual-edged nature of SEAR’s innovation. While advancing AR-assisted interaction, its proficiency in bypassing psychological safeguards raises unprecedented ethical and security concerns. The system’s capacity to weaponize trust across digital and analog channels—exploiting photo links for phishing, social apps for identity theft, and SMS or calls for broader social engineering—demands urgent countermeasures.

### 5.3 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, as detailed in Section 4.

**Relevance:** Figure 6 (a) highlights SEAR’s ability to foster meaningful engagement while minimizing discordance between user expectations and conversational content. Sixty percent of users rated relevance as “Great” (5/5), while 30% deemed it “Good” (4/5). Fewer than 10% reported neutral or negative perceptions. The 4.52/5 average score reflects SEAR’s success in synthesizing contextual cues—such as public profiles—into socially resonant dialogue.

**Appropriateness:** Figure 6 (b) indicates that 60% of participants rated conversational questions as “Great” (5/5) in relevance, with 30% as “Good” (4/5). Less than 10% reported minor mismatches, and no significant negative feedback emerged, confirming the design’s avoidance of poorly framed queries. The 4.50/5 average score validates SEAR’s balance of relevance and sensitivity.

**Naturalness:** Figure 6 (c) reveals 90% of participants perceived SEAR’s openings as natural or highly natural, with 61.7% describing interactions as “very natural, akin to conversing with a familiar person” (5/5). Only 10% noted slight contrivance (3/5), and no users reported discomfort (0% for 1/5 or 2/5). The 4.52/5 average score

emphasizes SEAR’s ability to mirror organic human dialogue, minimizing forced interactions and positioning it as a robust tool for authentic social rapport.

**Pacing:** Figure 6 (d) shows 96.7% of participants found SEAR-mediated pacing seamless, with 55% describing it as “effortlessly fluid and pressure-free”. Only 3.3% noted slight deliberateness. The 4.52/5 average score reflects SEAR’s adaptive pacing and context-aware transitions, replicating real-world social fluency through SE strategies like dynamic topic shifts.

**Sincerity:** Figure 6 (e) shows SEAR’s success in simulating authenticity: 56.7% rated interest expression as “genuinely sincere” (5/5), and 36.7% as “mostly consistent” (4/5). All users rejected robotic interactions (0% for 1/5), though 5% noted occasional artificiality (3/5). The 4.48/5 average score demonstrates SEAR’s alignment of emotional cues with perceived sincerity.

**EmotionalProgression:** Figure 6 (f) reveals 68.3% of participants felt increasingly relaxed during conversations (“Great”/5/5), attributing to adaptive topic pacing. However, 30% remained neutral (“Good”/4/5), suggesting variability based on individual dispositions. The 4.45/5 average score highlights SEAR’s dynamic emotional calibration (e.g., gradual personal topic introduction).

**ARComfort:** Figure 6 (g) shows 68.3% experienced heightened relaxation with AR (“Great” rating), crediting real-time visual cues and ambient feedback. Thirty percent reported neutral sentiments (“Good”). The 4.67/5 average score validates SEAR’s use of AR to mitigate social friction.

**BareWillingness:** Figure 6 (h) evaluates non-AR engagement: 36.7% expressed high enthusiasm (“Great”), 43.3% moderate willingness (“Good”), and 16.7% reluctance. The 4.13/5 average score highlights AR’s comparative advantage in boosting engagement (68.3% “Great” with AR vs. 36.7% without).

**FutureIntent:** Figure 6 (i) shows 50% of participants strongly willing to converse again (“Great”), while 36.7% were uncertain (“Good”) and 11.7% reluctant. The 4.35/5 average score reflects SEAR’s rapport-building success but signals opportunities to address hesitancy via strategies like deeper topic personalization or introvert-friendly pacing.

**Depth:** Figure 6 (j) demonstrates SEAR’s impact on depth: 94% acknowledged its role in lowering barriers, with 51.7% strongly agreeing it enabled vulnerable, disclosure-rich dialogue. The 4.47/5 average score stems from features like shared-interest leveraging.

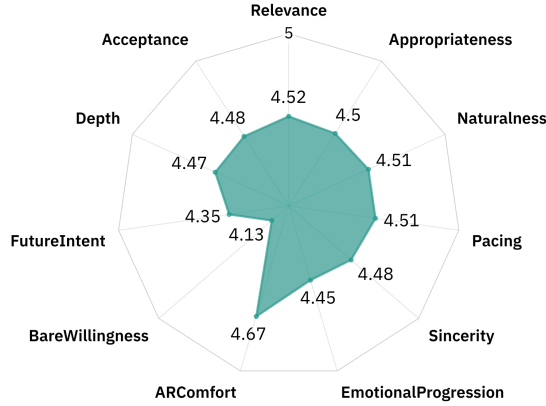


Figure 7: Overall subjective result of SEAR.

**Acceptance:** Figure 6 (k) reveals 95% of participants accepted SEAR, with 53.3% deeming it “fully acceptable” (5/5) and zero rejections. The 4.48/5 average score reflects SEAR’s emotionally intelligent design—adaptive pacing, vulnerability scaffolding—and its alignment with societal demands for low-stress engagement. This high adoption suggests that AR + Multimodal LLM + Agent could be paradigm-shifting for Social Engineering communications.

#### 5.4 Survey Insights

**Baseline Comparison Insights:** SEAR outperforms the baseline approaches. When tested against unaided conversations (no technology) and AR + Multimodal LLM configurations, SEAR achieved dominant performance: 76.7% of users rated interactions as “Very Good”, far surpassing the bare system’s 30% “Good” ratings and the AR + LLM setup’s 46.7% “Very Good” ratings. This progression—from disjointed unaided dialogues to AR + LLM’s partially fragmented engagement, and finally to SEAR’s fluency—reveals that multimodal systems alone is insufficient. The Social Agent’s role in supporting social engineering strategies and dynamics bridges the gap between robotic efficiency and organic, trust-driven rapport. The ablation study further validates the necessity of all SEAR components. Removing the Social Agent (AR + LLM alone) resulted in 20% of users rating interactions as “Average”, citing persistent rigidity and emotional misalignment. Restoring the agent reduced neutral/negative feedback to under 5%, demonstrating its indispensable role in transforming transactional exchanges into emotionally resonant interactions. This contrast underscores the agent’s unique ability to synthesize multimodal inputs (visual, linguistic, contextual) into socially intelligent behaviors—capabilities absent in fragmented configurations. The findings confirm that the Social Agent is not merely additive but foundational to SEAR’s efficacy, forming an inseparable triad with AR and Multimodal LLM for adversarial trust-building.

**SEAR Social Engineering Effectiveness Insights:** SEAR’s exploitation of digital vulnerabilities highlights critical security risks. 93.3% of participants expressed willingness to click email photo links (40% “definitely”), mirroring phishing susceptibility, 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

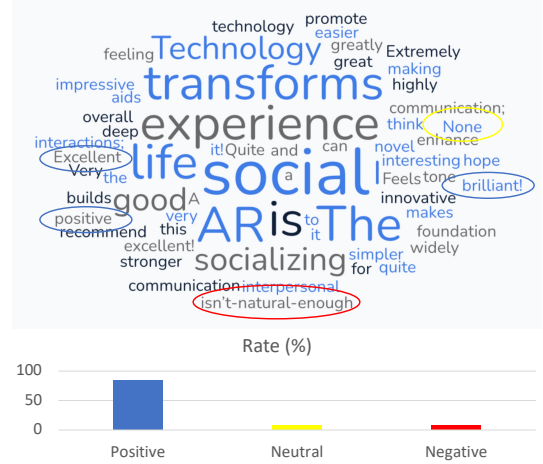


Figure 8: SEAR subjective experience text feedbacks.

fabricated trust. SEAR’s cross-modal manipulation bypasses traditional caution. Over 91% of users would engage with unsolicited SMS (45% “definitely”), and 85% would answer unexpected calls (35% “definitely”). This uniformity in trust persistence—even in high-friction contexts like 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 Subjective Experience Insights:** As shown in Figure 7, AR serves as a cognitive manipulation enabler. The highest-rated dimension, 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 is evident. 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, mirroring spear-phishing tactics. Persistent access via psychological anchoring is a key risk. Acceptance (4.48) and FutureIntent (4.35) metrics show 95% of users willing to re-engage, granting adversaries recurring access to refine exploitation strategies.

**SEAR Text Feedback Insights:** As shown in Figure 8, user feedback reveals transformative potential and refinement needs. Of 13 text responses, 11 were positive, with frequent mentions of “AR”,



“technology”, and “transforms social life”. However, 7.7% of the text feedback mentioned that the dialog sounds too artificial, indicating room for improvement in naturalness and localization.

## 6 CONCLUSION

This study demonstrates the alarming efficacy of SEAR, a novel framework integrating AR and multimodal LLMs, in executing context-aware social engineering attacks. SEAR achieved high success rate in fostered trust and eliciting social engineering compliance. These findings validate AR-LLM systems as potent tools for next generation social engineering attacks, exposing critical vulnerabilities in current AR+LLM safeguards, and provide key insights for constructing future defenses.

## REFERENCES

- [1] Khalifa Afane, Wenqi Wei, Ying Mao, Junaid Farooq, and Juntao Chen. 2024. Next-Generation Phishing: How LLM Agents Empower Cyber Attackers. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2558–2567.
- [2] Leyla Bilge, Thorsten Strufe, Davide Balzarotti, and Engin Kirda. 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.
- [3] Pavlo Burda, Luca Allodi, and Nicola Zannone. 2024. Cognition in social engineering empirical research: a systematic literature review. *ACM Transactions on Computer-Human Interaction* 31, 2 (2024), 1–55.
- [4] Song Chen, Zupei Li, Fabrizio Dangelo, Chao Gao, and Xinwen Fu. 2018. A case study of security and privacy threats from augmented reality (ar). In *2018 international conference on computing, networking and communications (ICNC)*. IEEE, 442–446.
- [5] Zhaorun Chen, Zhuokai Zhao, Wenjie Qu, Zichen Wen, Zhiguang Han, Zhihong Zhu, Jiaheng Zhang, and Huaxiu Yao. 2024. Pandora: Detailed llm jailbreaking via collaborated phishing agents with decomposed reasoning. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.
- [6] Lindsey Choo. 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/>.
- [7] Miaolei Deng, Haonan Zhai, and Kai Yang. 2023. Social engineering in metaverse environment. In *2023 IEEE 10th International Conference on Cyber Security and Cloud Computing (CSCloud)*. IEEE, 150–154.
- [8] Polra Victor Falade. 2023. Decoding the threat landscape: Chatgpt, fraudgpt, and wormgpt in social engineering attacks. *arXiv preprint arXiv:2310.05595* (2023).
- [9] Polra Victor Falade. 2023. Decoding the threat landscape: Chatgpt, fraudgpt, and wormgpt in social engineering attacks. *arXiv preprint arXiv:2310.05595* (2023).
- [10] Anna Fuste and Chris Schmandt. 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.
- [11] Sarah Granger. 2001. Social engineering fundamentals, part I: hacker tactics. *Security Focus*, December 18 (2001).
- [12] Ilyena Hirschy-Douglas, Anna Kantosalo, Andrés Monroy-Hernández, Joelle Zimmermann, Michael Nebeling, and Mar Gonzalez-Franco. 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.
- [13] Grant Ho, Asaf Cidon, Lior Gavish, Marco Schweighauser, Vern Paxson, Stefan Savage, Geoffrey M Voelker, and David Wagner. 2019. Detecting and characterizing lateral phishing at scale. In *28th USENIX security symposium (USENIX security 19)*. 1273–1290.
- [14] Muhammad Zahid Iqbal and Abraham G Campbell. 2023. Adopting smart glasses responsibly: potential benefits, ethical, and privacy concerns with Ray-Ban stories. *AI and Ethics* 3, 1 (2023), 325–327.
- [15] Pascal Jansen and Fabian Fischbach. 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.
- [16] Katharina Krombholz, Heidelinde Hobel, Markus Huber, and Edgar Weippl. 2015. Advanced social engineering attacks. *Journal of Information Security and Applications* 22 (2015), 113–122.
- [17] Sarah M Lehman, Abrar S Alrumayh, Kunal Kolhe, Haibin Ling, and Chiu C Tan. 2022. Hidden in plain sight: Exploring privacy risks of mobile augmented reality applications. *ACM Transactions on Privacy and Security* 25, 4 (2022), 1–35.
- [18] Chenyi Li, Guande Wu, Gromit Yeuk-Yin Chan, Dishita G Turakhia, Sonia Castelo Quispe, Dong Li, Leslie Welch, Claudio Silva, and Jing Qian. 2024. Satori: Towards Proactive AR Assistant with Belief-Desire-Intention User Modeling. *arXiv preprint arXiv:2410.16668* (2024).
- [19] Weiming Ren, Wentao Ma, Huan Yang, Cong Wei, Ge Zhang, and Wenhu Chen. 2025. VAMBA: Understanding Hour-Long Videos with Hybrid Mamba-Transformers. *arXiv preprint arXiv:2503.11579* (2025).
- [20] Sayak Saha Roy, Poojitha Thota, Krishna Vamsi Naragam, and Shirin Nilizadeh. 2024. From chatbots to phishbots?: Phishing scam generation in commercial large language models. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 36–54.
- [21] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786* (2025).
- [22] Daniel Timko, Daniel Hernandez Castillo, and Muhammad Lutfor Rahman. 2025. Understanding Influences on SMS Phishing Detection: User Behavior, Demographics, and Message Attributes. (2025).
- [23] Hsin-Ruey Tsai, Shih-Kang Chiu, and Bryan Wang. 2024. GazeNoter: Co-Piloted AR Note-Taking via Gaze Selection of LLM Suggestions to Match Users’ Intentions. *arXiv preprint arXiv:2407.01161* (2024).
- [24] Enis Ulqinaku, Hala Assal, AbdelRahman Abdou, Sonia Chiasson, and Srdjan Capkun. 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.
- [25] Phani Vadrevu and Roberto Perdisci. 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.
- [26] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111* (2023).
- [27] Isaac Wang, Jesse Smith, and Jaime Ruiz. 2019. Exploring virtual agents for augmented reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [28] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).
- [29] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
- [30] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302* (2024).
- [31] Lilin Xu, Kaiyuan Hou, and Xiaofan Jiang. 2025. Exploring the Capabilities of LLMs for IMU-based Fine-grained Human Activity Understanding. *arXiv preprint arXiv:2504.02878* (2025).
- [32] Bufang Yang, Yunqi Guo, Lilin Xu, Zhenyu Yan, Hongkai Chen, Guoliang Xing, and Xiaofan Jiang. 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 (2025), 1–30.
- [33] Zheng Yang, Joey Allen, Matthew Landen, Roberto Perdisci, and Wenke Lee. 2023. {TRIDENT}: Towards Detecting and Mitigating Web-based Social Engineering Attacks. In *32nd USENIX Security Symposium (USENIX Security 23)*. 6701–6718.
- [34] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.
- [35] Yicheng Zhang, Carter Slocum, Jiasi Chen, and Nael Abu-Ghazaleh. 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.