SocialMind: LLM-based Proactive AR Social Assistive System with Human-like Perception for In-situ Live Interactions

BUFANG YANG[†], The Chinese University of Hong Kong, China
YUNQI GUO[†], The Chinese University of Hong Kong, China
LILIN XU[†¶], Columbia University, United States
ZHENYU YAN[‡], The Chinese University of Hong Kong, China
HONGKAI CHEN, The Chinese University of Hong Kong, China
GUOLIANG XING, The Chinese University of Hong Kong, China
XIAOFAN JIANG, Columbia University, United States

Social interactions are fundamental to human life. The recent emergence of large language models (LLMs)-based virtual assistants has demonstrated their potential to revolutionize human interactions and lifestyles. However, existing assistive systems mainly provide reactive services to individual users, rather than offering in-situ assistance during live social interactions with conversational partners. In this study, we introduce SocialMind, the first LLM-based proactive AR social assistive system that provides users with in-situ social assistance. SocialMind employs human-like perception leveraging multi-modal sensors to extract both verbal and nonverbal cues, social factors, and implicit personas, incorporating these social cues into LLM reasoning for social suggestion generation. Additionally, SocialMind employs a multi-tier collaborative generation strategy and proactive update mechanism to display social suggestions on Augmented Reality (AR) glasses, ensuring that suggestions are timely provided to users without disrupting the natural flow of conversation. Evaluations on three public datasets and a user study with 20 participants show that SocialMind achieves 38.3% higher engagement compared to baselines, and 95% of participants are willing to use SocialMind in their live social interactions.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**.

Additional Key Words and Phrases: LLMs, Augmented Reality, AR Glasses, Multi-modal Sensor Data, Proactive Assistive Systems, Social Interaction, Internet of Things

ACM Reference Format:
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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 1, Article 23 (March 2025), 30 pages. https://doi.org/10.1145/3712286

[†]Co-primary authors.
[¶]Part of this work was completed while Lilin Xu was visiting the AIoT Lab at The Chinese University of Hong Kong.
[‡]Corresponding author.

Authors' Contact Information: Bufang Yang, The Chinese University of Hong Kong, China, bfyang@link.cuhk.edu.hk; Yunqi Guo, The Chinese University of Hong Kong, China, yunqiguo@cuhk.edu.hk; Lilin Xu, Columbia University, United States, lx2331@columbia.edu; Zhenyu Yan, The Chinese University of Hong Kong, China, zyyan@ie.cuhk.edu.hk; Hongkai Chen, The Chinese University of Hong Kong, China, hkchen@ie.cuhk.edu.hk; Guoliang Xing, The Chinese University of Hong Kong, China, glxing@cuhk.edu.hk; Xiaofan Jiang, Columbia University, United States, jiang@ee.columbia.edu.

ACM 2474-9567/2025/3-ART23
https://doi.org/10.1145/3712286

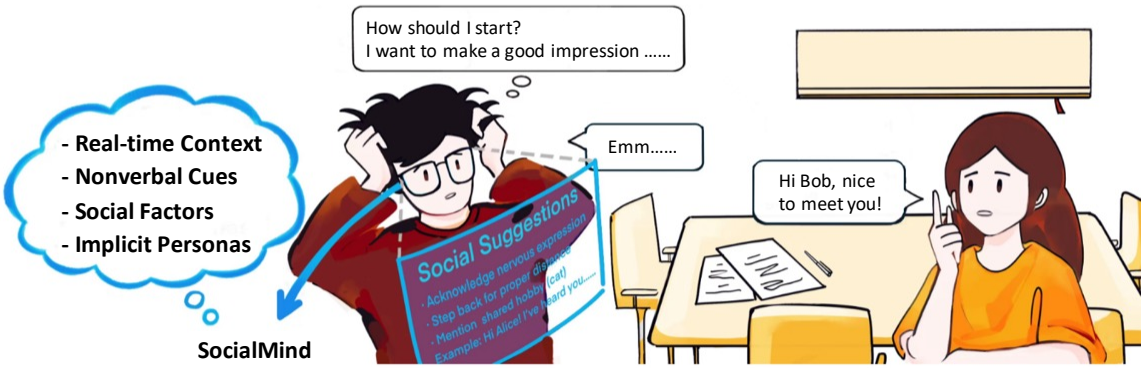


Fig. 1. Overview of SocialMind. SocialMind provides in-situ social assistance to the user to help the user during live social interactions with conversational partners. SocialMind automatically performs human-like social perception, generates social suggestions to assist users, and proactively displays them on the user's AR glasses as the conversation proceeds. Users can seamlessly refer to these suggestions while interacting with conversational partners.

1 Introduction

Social interactions are a crucial determinant of quality of life, significantly impacting both physical and mental health by enhancing communication skills and alleviating stress [3]. However, over 15 million individuals in America experience social anxiety when anticipating or engaging in social interactions [5]. Even people without social anxiety may feel anxious when interacting with certain individuals, such as senior managers and unfamiliar colleagues, which can reduce their overall well-being. With the surge of Large Language Models (LLMs) and their reasoning capabilities [7, 67], numerous LLM-based personal virtual assistants have been developed to enhance individuals' overall well-being. However, existing LLM-based personal virtual assistants, such as writing assistants [26, 60], fitness assistants [68, 78], and coding assistants [22], focus solely on serving individual users rather than supporting live social interactions with conversational partners.

In addition to general virtual assistants, LLM-based social assistive systems have been proposed recently, aiming to support autistic patients [41], provide knowledge consultation on social etiquette [6], and resolve cultural conflicts in communication [37, 81, 83]. These assistive systems either function in a reactive "query-response" manner to address users' explicit social-related questions [6, 41], or act as post-processing modules to detect and remediate norm violations in social conversations [36, 37], rather than providing in-situ assistance during live social interactions with conversational partners. However, human face-to-face social interactions are complex behaviors that necessitate considering various types of information, including verbal, and non-verbal behaviors, social environment, social purpose, and the personal backgrounds of both parties [34]. Therefore, assistive systems in live social interactions must act like humans to perceive diverse information, such as the context of the live social conversation, the multi-modal nonverbal behaviors of other parties [21], and various social factors [82]. They should also incorporate this information for reasoning and dynamically adjust their strategies for providing instant social suggestions to the user. This highlights a research gap in providing proactive social assistance during live, face-to-face social interactions involving conversational partners.

To address this research gap, we propose a proactive social assistive system for live social interactions leveraging LLMs and the multi-modal sensor data from Augmented Reality (AR) glasses. However, several unique challenges remain in developing such a system. First, *an assistive system for live social interactions requires providing users with instant and in-situ assistance during natural conversations with other parties.* Unlike existing personal assistants and assistive systems designed for single users [26, 53, 68, 75], live social interactions involve conversational partners, presenting additional challenges in providing instant responses to the user without disrupting the natural flow

of conversation. Second, *nonverbal behaviors are essential for social communication, yet they are challenging for LLMs to comprehend.* LLMs are trained exclusively on text corpora, whereas nonverbal behaviors such as facial expressions, gestures, and physical proximity involve multi-modal information [21], posing challenges for LLMs in understanding and integrating these cues to generate nonverbal cue-aware social suggestions. Third, social suggestions need to consider the implicit personal backgrounds and interests of both parties to enhance their engagement. However, *natural social conversations often lack explicit queries for knowledge retrieval, posing challenges for system personalization.* How to integrate implicit personas like personal interests and background into social suggestions remains challenge.

In this paper, we introduce SocialMind, the first LLM-based proactive AR social assistive system that provides users with in-situ social assistance in live social interactions with other parties. Figure 1 shows the overview of SocialMind. SocialMind leverages the multi-modal sensors on AR glasses to perform human-like perception, including verbal and nonverbal cues and social factor information. Additionally, SocialMind extracts the implicit personas of both parties through social interactions. Then, SocialMind integrates these cues and utilizes a multi-tier collaborative social suggestion generation strategy that incorporates a cache with social-factor priors and an intention infer-based reasoning approach. This strategy enables SocialMind to provide timely, in-situ social assistance to the user during natural conversations with partners. Through a proactive response mechanism, social suggestions are displayed on AR glasses, enabling the user to seamlessly refer to them while interacting with their conversational partner. We summarize the contributions of this paper as follows.

- We introduce SocialMind, the first LLM-based proactive AR social assistive system providing users with **in-situ assistance** during live social interactions. We develop a multi-tier collaborative suggestion generation strategy, incorporating a social factor-aware cache and intention infer-based reasoning, along with a proactive update mechanism. This ensures users receive timely and in-situ social suggestions, which are displayed on AR glasses without disrupting the natural flow of conversation.
- We design a human-like perception mechanism that enables SocialMind to automatically leverage multi-modal sensor data to perceive social cues, and develop a multi-source social knowledge reasoning approach to incorporate these cues into LLM reasoning, dynamically adjusting strategies for social assistance.
- We develop an implicit persona adaptation approach that enables SocialMind to generate customized social suggestions, enhancing the engagement of both parties in live social interactions.
- Motivated by a user survey involving 60 participants to understand their social experiences and preferences for social assistance, we designed and implemented SocialMind on AR glasses and validated its effectiveness using three public datasets and real-world tests. Evaluations on these datasets and a user study with 20 participants revealed that SocialMind achieves a 38.3% higher engagement compared to baselines, with 95% of participants expressing a willingness to use SocialMind in live social interactions.

2 Related Work

2.1 LLM-based Personal Assistants

Voice assistants are widely used in daily lives on various commercial mobile devices, such as Apple's Siri [1] and Google Assistant [30]. Recently, LLM-based virtual assistants have been developed, such as fitness assistants [68, 77, 78], writing assistants [26, 60], and coding assistants [22]. OS-1 [75] is a virtual companion on smart glasses offering companionship by recording daily activities and chatting with users. UbiPhysio [68] is a fitness assistant that provides natural language feedback for daily fitness and rehabilitation exercises, improving workout quality. Moreover, recent studies develop personal assistants for older adults [27, 79] and individuals with impairments [44]. EasyAsk [27] is a search assistant for older adults, accepting both audio and text inputs to provide app tutorials based on their queries. Talk2Care [79] is a voice assistant designed to engage in conversations with older adults to gather health information for healthcare providers. Additionally, studies like PEARL [60] and

Table 1. A summary of the recent LLM-based applications in social and communication (● means included).

Approach	Base LLM	Social Assistance	Multi Party Interactions	Multi-modal Sensor Data	Persona-lization	Interactive Mode	System Settings
Social-LLM [43]	Sentence-BERT	○	○	○	○	Reactive	PC
Paprika [41]	GPT-4	●	○	○	○	Reactive	PC
Tianji [6]	InternLM	●	○	○	○	Reactive	PC
Hua *et. al* [37]	GPT 3.5, Atom-7B-Chat	●	●	○	○	Reactive	PC
SADAS [36]	ChatGPT	●	●	○	○	Reactive	HoloLens
OS-1 [75]	GPT4, Gemini, Llama2	○	○	●	●	Proactive	Glasses
PRELUDE [26]	GPT-4	○	○	○	●	Reactive	PC
SocialMind	GPT-4o, Llama-3.1	●	●	●	●	Proactive	Glasses

PRELUDE [26] develop LLM-based writing assistants that adapt outputs to user preferences using retrieval augmentation [60] or interactive learning [26]. However, these systems focus solely on single-user human-to-computer interactions, considering only the user's unilateral goals and inputs. SocialMind takes a further step by providing users with social assistance during live, face-to-face interactions involving other parties.

2.2 Social Assistive Systems

Pre-LLM Era. SocioGlass [74] builds a biography database, using smart glasses and facial recognition to retrieve profiles with background and interests for social interaction assistance. Another study explores the use of smart glasses to support social skills learning in individuals with autism spectrum disorder [45]. However, these systems are limited to displaying social skills or biographies on-screen, lacking the context of real-time social conversation.
LLMs for Social Assistance. Paprika [41] employs LLMs to provide social advice to autistic adults in the workplace. Results show that autistic workers prefer interactions with LLMs, demonstrating LLMs' potential to offer social advice. Tianji [6] is an LLM that comprehends social dynamics, offering social skill guidance by answering questions, like how to resolve conflicts. Social-LLM [43] integrates users' profiles and interaction data to generate user embeddings for user detection. However, these works are reactive conversational systems limited to social Q&A or user behavior prediction, rather than providing instant social assistance when users are interacting with others. Some studies also explore the impact of social norms and their violations in communication and negotiation, using simulations with multiple LLM agents [36, 37, 81, 83]. SADAS [36] is a dialogue assistant that checks user input for social norm violations to improve cross-cultural communication. Kim *et. al* [46] develops a dialogue model to detect unsafe content and generate prosocial responses. However, these systems provide post-assistance, addressing social norm violations in user text-only input only after it has been entered. SocialMind focuses on live face-to-face scenarios, proactively perceiving multi-modal nonverbal cues and conversation context to provide instant social suggestions, enabling users to refer to them before speaking.

2.3 Proactive Conversational Systems

Reactive conversational systems follow the "receive and respond" paradigm, exemplified by writing assistants [26, 60] and coding assistants [22], which generate an answer based on the user's input, without further interaction. Proactive conversational systems can initiate and steer conversations through multi-turn interactions with users [19]. OS-1 [75] utilizes personal daily logs, historical context, and perceived environmental information to proactively engage users, serving as a virtual companion. DrHouse [78] is a proactive multi-turn diagnostic system that uses expert medical knowledge and sensor data for multi-turn assessments. WorkFit [8] is a proactive voice assistant that detects sedentary behavior in workers and generates voice interventions and health suggestions. However, existing proactive conversational systems are limited to individual user scenarios. There remains a gap in research on proactive assistive systems for live social interactions involving conversational partners.

2.4 LLM Personalization and Acceleration

LLM Caching. Caching solutions have been utilized in LLM reasoning systems to reduce repetitive computations, including caching LLM response and caching intermediate states [11, 25, 29, 49, 80]. GPT-cache [11] and SCALM [49] employ semantic cache to store the LLMs responses. Additionally, numerous studies employ key-value (KV) cache, reusing attention states during LLM response generation, to reduce inference costs [25, 29, 80]. CachedAttention [25] reuses the KV cache of historical tokens in multi-turn conversations. Prompt Cache [29] resues the attention states of the overlapped text segments among different prompts. Unlike general cache designs, SocialMind incorporates social factor priors into the cache to enhance accuracy.

Streaming and Real-time LLMs. Real-time AI assistants have been developed recently, such as Gemini Live [2]. It supports users to interrupt conversations and assists with daily tasks on mobile phones. Additionally, some studies explore the real-time speech LLMs [52, 64, 71]. Mini-Omni [71] integrates hearing, speaking, and thinking into speech foundation models for real-time conversation. Speech ReaLLM [64] achieves real-time speech recognition by streaming speech tokens into LLMs for reasoning without waiting for the entire utterance or changing the LLM architecture. However, these systems focus on general speech recognition and lack the integration of multi-modal social knowledge, limiting their utility in live social interactions. SocialMind is designed to proactively provide social suggestions during live interactions involving multiple participants.

3 A Survey on Social Assistance Needs

To understand the demand for social assistants during interactions, we conduct a survey exploring user experience, preferences, and needs regarding live social interactions. The results and findings guide the design of our system.

3.1 Design of Questionnaire

The questionnaire comprises three sections, totaling 14 questions. The questions are summarized as follows:

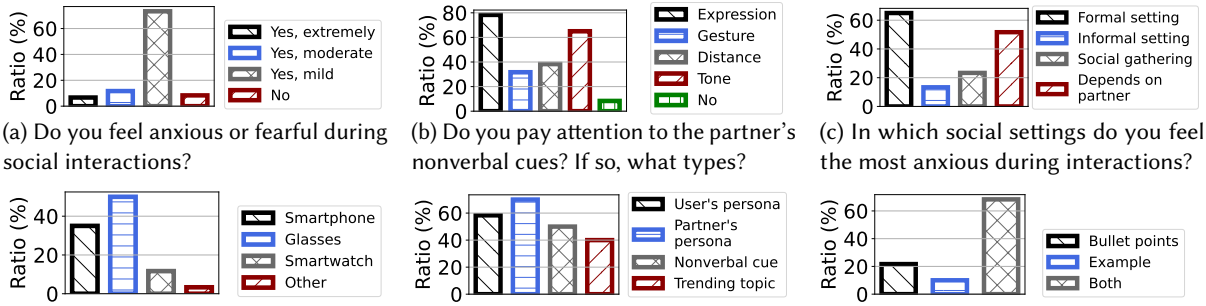
- ***P1:*** This section is designed to capture participants' social experiences, including their experiences of social awkwardness, awkwardness sources, and attention to nonverbal behaviors during social interactions.
- ***P2:*** The second section assesses the needs and preferences for virtual social assistance technologies. It includes questions about participants' attitudes toward social assistance during live interaction, preferred devices, desired social situations, desired content of suggestion, and assistive information format. It also examines participants' preferred information display methods and tolerance for system latency.
- ***P3:*** The final section explores participants' attitudes toward privacy and comfort in the context of virtual social assistance technologies, assessing their willingness to interact with users utilizing such assistants and concerns about potential personal data capture during interactions.

We collect 60 questionnaires in total, and summarize the results and findings as follows.

3.2 Social Experience and Awkwardness

Among the participants, 18.3% consider themselves to enjoy interacting with others, while the remaining participants describe themselves as not enjoying it as much. Besides, only less than 10.0% of the participants report being completely at ease during social interactions. As shown in Figure 2a, 91.7% claim that they experience some level of awkwardness in social situations, indicating that social awkwardness is pretty common in daily life.

The survey results indicate social awkwardness comes from various sources. Specifically, more than 60.0% report experiencing awkwardness when interacting with workplace superiors or professors. This trend extends to formal events like meetings. Peer interactions contribute as well, with 40.0% feeling nervous when interacting with colleagues or fellow students, particularly in initial encounters. Furthermore, 31.7% report awkwardness when interacting with long-lost acquaintances, and nearly half feel anxious when communicating with unfamiliar relatives. Moreover, as Figure 2c shows, 65.0% experience stress in formal settings. Besides, over half also regard conversational partners as a key factor, indicating that personal relationships are vital in shaping social



(a) Do you feel anxious or fearful during social interactions?

(b) Do you pay attention to the partner's nonverbal cues? If so, what types?

(c) In which social settings do you feel the most anxious during interactions?

(d) For a virtual assistant, which device would you prefer to use it on?

(e) What kind of social advice would you like a virtual assistant to provide?

(f) How would you prefer the virtual assistant to provide help?

Fig. 2. Survey results for social experience and assistance demand.

awkwardness. These results indicate that social awkwardness is most pronounced in situations involving authority figures, formal settings, and unfamiliarity.

Furthermore, the results reveal that nonverbal social behaviors play an important role in social interactions, particularly facial expressions, tone of voice, personal distance, and gestures. Figure 2b shows that only 8.3% overlook nonverbal behaviors while the majority consider them essential during interactions. Specifically, facial expressions are noted by nearly 80.0% of the participants, and tone of voice is noted by 65.0%. Besides, 38.3% are attentive to personal distance, and 31.7% regard gestures as supplementary cues. Despite these nonverbal behaviors' significance, their indirect nature presents challenges, suggesting a need for support in interpreting nonverbal cues.

3.3 The Demand and Preference for a Social Assistant

The questionnaire's second section reveals that the preference for a virtual social assistant aligns with the social awkwardness experienced by participants. Notably, 70.0% believe that a virtual assistant offering instant social suggestions during interactions would be beneficial, indicating a clear demand for social assistance technology.

Individuals desire assistance during social interactions in certain scenarios. To be specific, 66.7% need assistance when feeling uncertain or embarrassed about what to say, 56.7% when interacting with specific individuals, particularly authority figures, and nearly half when unsure how to respond or initiate the conversation with a goal in mind. Furthermore, participants also have content preferences for a virtual assistant's instant suggestions. Specifically, participants value both information on conversational partners' and their own interests and backgrounds, with 70.0% preferring insights into partners' personas and over 50.0% interested in their own profiles. Additionally, 40.0% seek updates on trending topics, and half want social cues about nonverbal behaviors. These results suggest that an effective virtual assistant should offer social assistance with human-like perception.

The results in Figure 2d show that over half of the participants prefer glasses as assistive devices since glasses are convenient and appear natural in conversation. Furthermore, for information display, 93.3% prefer text projected in their field of vision. For assistive information format, as demonstrated in Figure 2f, 68.3% prefer both summarized bullet points and example sentences, indicating a need for concise and direct suggestions. Moreover, instant assistance is preferred with 90.0% emphasizing instant delivery. These results suggest a potential demand for employing AR glasses to provide in-suit social assistance, offering instant, easily accessible information without disrupting the conversation flow.

3.4 Privacy and User Comfort

Privacy and user comfort are critical factors in the adoption and acceptance of virtual social assistance technologies. Results reveal strong openness to such technologies, with 88.3% willing to engage with users employing such

assistants. However, when confronted with specific privacy concerns, such as image capture during interactions, user comfort levels decrease. Despite this, 63.3% are willing to continue conversations. This indicates that while privacy concerns are present, they do not significantly deter interest and demand for social assistance technologies, highlighting a generally positive reception.

3.5 Findings Summary

We summarize our key findings as follows:

- Social awkwardness is pretty common in daily life, particularly in interactions with authority figures, formal settings, and unfamiliar situations. This reveals the potential benefits of virtual social assistance.
- Nonverbal behaviors like gestures, facial expressions, and personal distance are essential in interactions, as people naturally perceive and focus on these cues during conversations. An effective virtual assistant should therefore provide assistance with human-like perception for nonverbal cues.
- Participants show strong interest in a virtual social assistant that offers instant guidance to reduce social awkwardness. They prefer assistance in specific scenarios, certain suggestion content, natural integration via glasses, as well as concise and instant suggestions. These results indicate a clear demand for a proactive system based on AR glasses to provide effective social assistance during live interactions.

These findings further motivate the design of our proactive social assistive system for in-situ live interactions based on AR glasses and LLMs.

4 System Design

4.1 System Overview

SocialMind is an LLM-based proactive AR social assistive system capable of human-like perception, providing users with in-situ assistance during live social interactions. Figure 3 shows the system overview of SocialMind. SocialMind first leverages the multi-modal sensor data, including audio, video, and head motion, to achieve human-like perception in social contexts (§ 4.2). It automatically extracts nonverbal and verbal behaviors, and parses social factor cues. Meanwhile, SocialMind identifies implicit personas from social contexts and performs implicit persona adaptation (§ 4.3). The extracted verbal and nonverbal behaviors, social factors, and implicit persona cues are then integrated into the LLMs for reasoning (§ 4.4). Finally, SocialMind employs a multi-tier collaborative reasoning strategy with a social factor-aware cache and intention infer-based reasoning approach to generate in-situ social suggestions (§ 4.5). These suggestions are displayed on AR glasses through a proactive response mechanism to assist users in live social interactions without disrupting the natural flow of conversations.

We chose AR glasses for social assistance over devices like smartphones or smartwatches for three main reasons. First, AR glasses for daily wear are increasingly accepted, as seen in applications like captioning and translation [33, 39, 62, 65]. Second, AR glasses offer a non-distracting, hands-free solution that enables users to maintain eye contact during social interactions without interrupting the natural flow of conversation. Finally, our survey indicates that most participants favor glasses as the ideal hardware for embedding a social assistive system in live interactions over other devices.

4.2 Human-like Perception in Social Context

Existing studies on social assistive systems focus solely on single-user human-to-computer interactions and follow a reactive paradigm, conducting either question-answering [6, 41] or remediating cultural violations [37, 81, 83]. However, live social interactions involve multi-modal cues such as nonverbal behaviors and social factors, posing challenges to existing text-only LLMs in providing comprehensive social suggestions. SocialMind employs a human-like perception approach that can leverage the multi-modal sensor data to extract social cues during live social interactions.

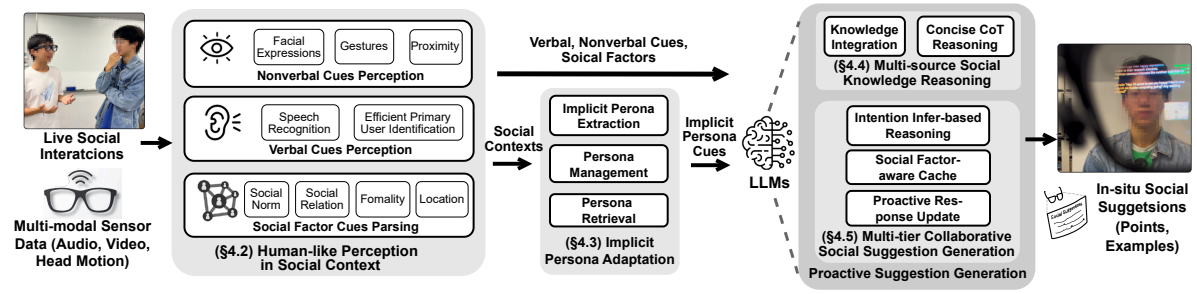


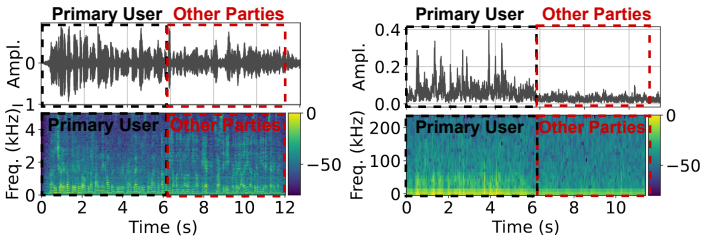
Fig. 3. System overview of SocialMind. SocialMind leverages the multi-modal sensor data to achieve human-like perception. The extracted verbal and nonverbal cues, social factors, and implicit personas are integrated into LLMs, generating in-situ social suggestions with points and examples displayed on the user's AR glasses.

4.2.1 Nonverbal Cues Perception. Nonverbal behaviors play a crucial role in face-to-face social interactions [21]. For example, facial expressions like confusion and frowning can indicate a person's emotional state during face-to-face social conversations [24]. Additionally, gestures can reveal a person's implicit perspectives, such as their understanding, intentions, or agreement, during social interaction.

SocialMind proactively perceives the nonverbal behaviors of the conversational partners and leverages these implicit cues to adjust social strategies and assist the user. However, nonverbal behaviors such as facial expressions, gestures, and physical proximity are captured by multi-modal sensors. Directly offloading the raw multi-modal data to the cloud server incurs significant bandwidth usage, high latency, and raises privacy concerns. To address these challenges, SocialMind employs multiple lightweight yet specialized small models on AR glasses to efficiently process raw data locally. Specifically, we first employ MediaPipe Holistic [31] in SocialMind to generate human poses, including facial mesh and hand poses. These facial mesh and hand poses are then further processed by different specialized models to generate nonverbal cues (§ 5.1.1). Finally, these nonverbal cues are incorporated into the LLMs to generate nonverbal cues-aware social suggestions (§ 4.4). Table 3 shows the details of the nonverbal cues detected in SocialMind, including facial expressions, gestures, and physical proximity [55]. We selected these nonverbal cues based on feedback from our user survey in § 3 and because existing studies indicate that they are the most representative forms of nonverbal communication during face-to-face social interactions [21].

4.2.2 Efficient Primary User Identification. Since SocialMind focuses on live face-to-face social interactions with conversational partners, it requires efficient and robust identification of the primary user and other participants. In recent years, user identification technology has been extensively developed through the exploration of diverse personal behavioral traits as biometric identifiers [16, 17, 28, 72]. Voice fingerprinting [28] can be used for speaker identification, but it introduces additional overhead from registration and raises security concerns, such as voice synthesis and replay attacks [48]. This is evidenced by Microsoft's recent closure of its speaker recognition service [58]. Volume-based solutions [17] utilize low-frequency energy to differentiate the primary user's speech from that of nearby individuals, but their robustness is limited by environmental noise and variations in the user's speaking volume. To address these challenges, SocialMind leverages a lightweight primary user identification approach leveraging the vibration signals on the smart glasses as indicators.

We first conduct real-world measurements where the primary user wears smart glasses and engages in conversations with different partners. The smart glasses record the audio and vibration signals simultaneously. Figure 4 shows the waveform of the audio and vibration signals on the smart glasses during live social interactions. The primary user speaks during the first 6 seconds, while the conversational partner speaks during the last 6 seconds. Compared to the audio, the amplitude of the vibration exhibits a clear difference between the primary



(a) Audio signal as indicator. (b) Vibration signal as indicator.

Fig. 4. SocialMind's primary user detection.

Table 2. Details of Social Factors in SocialMind.

Social Factors	Sub-Categories
Social Norm	Greeting, Request, Apology, Persuasion
Social Relation	Peer-Peer, Elder-Junior, Mentor-Mentee, Student-Professor
Formality	Informal, Formal
Location	Office, Open Area, Restaurant, Conference Break

user's speaking period and the partner's speaking period. Therefore, we leverage the signal energy vibration signals as an indicator to detect the primary user. Specifically, we calculate the vibration signal's energy within the 3 ∼ 10 Hz range and use it as the indicator. Energies exceeding a certain threshold are detected as the primary user. We employ a grid search to determine the optimal threshold. The sample rate of the vibration signal in SocialMind is set to 466 Hz, which is significantly lower than the audio sample rate, thereby reducing bandwidth usage. Additionally, SocialMind transmits the vibration signal from the glasses to the server at regular intervals of 300 ms and sets the threshold for primary user detection at 1.1 on the server. Details on the threshold search and the overall detection performance of our approach compared to audio-based solutions can be found in § 5.3.3.

4.2.3 Social Factor Cues Parsing. Existing studies show that social factors play a vital role in social communication [14]. Social behaviors and speech content considered acceptable or unacceptable can vary significantly depending on different social factors such as social relation and formality [35]. For example, when making a request, the tone and content of our speech should vary significantly depending on whether we are addressing a familiar person or a superior, such as a professor or manager. Similarly, social norms differ between formal and informal occasions. Therefore, it is essential to incorporate these social factor cues into social suggestion generation strategies.

SocialMind leverages the social contexts to parse social factor cues. It supports two modes of social factor perception: reactive and proactive. In reactive perception mode, the social contexts are instructions provided by the user, describing their social goals, such as: "I am going to a social communication with a senior professor during a conference break, and my goal is to introduce my research work and establish a social connection with him." SocialMind utlizes LLMs with dedicated prompts to parse social factors from the user's instructions before initiating social interactions. If the user does not actively provide descriptions of social factors, SocialMind will operate in proactive mode to parse social factors. In such mode, the social contexts are the captured images with social environment information. Specifically, SocialMind pre-stores the images of various locations such as conferences, meeting rooms, and restaurants. SocialMind leverages the camera on the glasses to recognize the current location by mapping it with the pre-stored images, thereby generating location-based social factors. The social factors identified through either reactive or proactive modes will be used as a knowledge source and integrated into the LLMs for generating social suggestions (§ 4.4.2). Table 2 shows the social factors utilized in SocialMind, including social norm, social relation, formality, and location [82].

4.3 Implicit Persona Adaptation

Every individual has unique backgrounds, life experiences, and personal interests, which are abstracted into personas [10]. A social topic that connects the personal interests and backgrounds of both parties can enhance the engagement of both parties. An ideal social assistive system should proactively identify the implicit personas of both parties and incorporate these personas into the strategies for social suggestion generation. However, *natural social conversations often lack explicit queries to initiate the knowledge retrieval of personal historical databases, posing challenges for system personalization.* SocialMind employs an implicit persona adaptation approach to generate customized social suggestions, enhancing the engagement of both parties.

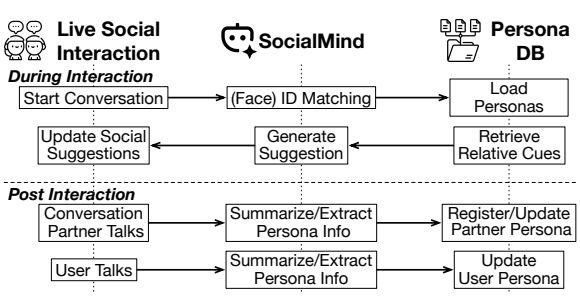


Fig. 5. Implicit persona adaptation in SocialMind.

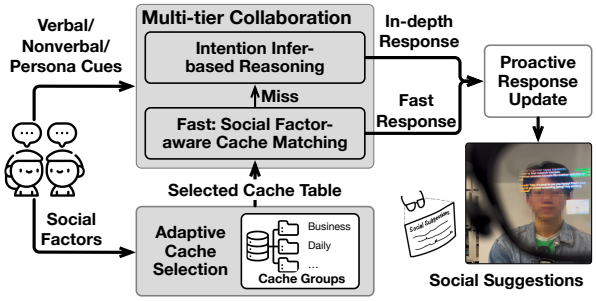


Fig. 6. Multi-tier collaborative social suggestion generation.

4.3.1 Implicit Persona Extraction. Existing personal assistant systems employ the user's explicit queries to retrieve historical data and provide personalized responses [18, 26, 78]. However, systems like writing assistants [26], emotional support assistants [18], and medical assistants [78] primarily function in a question-answering manner, relying on explicit queries to initiate the retrieval of the personal historical database. These explicit queries allow them to utilize the standard RAG techniques [47] to retrieve responses with high semantic similarity. However, natural social conversations lack explicit queries to initiate the retrieval of personal historical data, posing challenges in generating social suggestions that incorporate implicit personas.

To address this challenge, SocialMind employs an additional LLM to extract the implicit personas of both parties from historical conversations in advance, maintaining a *persona database*. Figure 5 shows the pipeline of the implicit persona adaptation in SocialMind. Specifically, the persona extraction occurs during the post-interaction phase, where an LLM extracts the implicit persona cues from the social conversations. These persona cues reflect the personal interests, experiences, and backgrounds of both parties. The persona database is organized according to individual identities, including those of the user and various conversational partners. Note that SocialMind will not engage in any privacy-infringing activities, such as actively crawling the social network data of other individuals based on facial recognition.

4.3.2 Persona Management. Since live experiences and personal interests evolve over time, SocialMind employs a persona management strategy to adapt to these emerging personas. Specifically, for new conversational partners, their persona cues will be directly registered in the persona database. For the user and previously met partners, SocialMind first utilizes LLMs to determine if any contradictory or similar cues already exist within the persona database for that particular identity. If no such cases are found, the incoming persona cues are registered and stored in the memory. If the incoming persona cues are semantically similar to existing ones, the two sets of cues are merged. Conversely, if the incoming persona cues are contradictory, the historical cues are removed and replaced by the new incoming ones.

4.3.3 Persona Retrieval. During live social interactions, SocialMind first performs persona retrieval using face ID matching to determine whether the conversational partners are in the database. This facial recognition is executed locally on the glasses to protect privacy. If the partners are found, the personas of both the user and the partner are loaded and integrated into LLMs as a knowledge source (§4.4). Otherwise, only the user's persona will be used. These implicit persona cues can help LLMs identify shared interests or experiences and generate customized social suggestions, thereby enhancing the engagement of both parties.

4.4 Multi-source Social Knowledge Reasoning

Existing social assistive systems primarily focus on text-based question answering and remediating social violations. SocialMind leverages multi-modal sensors to gather multi-source knowledge during live social interactions.

This section introduces the knowledge integration in SocialMind, which guides the LLMs to understand and utilize this knowledge in live social interactions, ensuring improved instruction-following performance and enhanced user Quality of Experience (QoE).

4.4.1 Knowledge Source. The knowledge source in SocialMind contains nonverbal cues (§ 4.2.1), the context of live social conversations (§ 4.2.2), social factors (§ 4.2.3), and implicit persona cues from both parties (§ 4.3.1). SocialMind also integrates external tools that provide the latest weather updates and trending social news, which are valuable sources for conversation topics. Since this information is updated daily and does not need to be retrieved during online interactions, it is pre-retrieved and incorporated into SocialMind's knowledge source.

4.4.2 Knowledge Integration. This subsection introduces the prompt used in SocialMind, which enables LLMs to integrate multi-source and multi-modal information for social suggestion generation. Specifically, the prompt in SocialMind consists of two parts: $prompt = prompt_{static} + prompt_{runtime}$, where $prompt_{static}$ represents the static portion and $prompt_{runtime}$ represents the runtime changing portion during live social interactions.

Static Prompt. This part of the prompt remains fixed and does not update any information during social interactions. It contains the overall instructions, task instructions, prior knowledge of nonverbal cues and their usage guidelines, as well as several few-shot demonstrations. Specifically, the overall instructions, together with the few-shot demonstrations, activate the LLM's capability to generate social suggestions. The task instructions enumerate various rules and requirements to enhance the LLM's ability to generate high-quality social suggestions. Additionally, we integrate literature and guidelines [21] on utilizing nonverbal cues during live social interactions as prior knowledge within the prompt. This includes categories of typical nonverbal cues and their corresponding coping strategies. For example, in the case of physical proximity, personal distance (1.5 to 4 feet) is common among family members or close friends and signifies the intimacy of the relationship [9]. This prior knowledge helps bridge the gap between the embedded knowledge of LLMs and the specific expertise required for effective nonverbal communication.

Runtime Prompt. This part of prompt dynamic changes during live social interactions. SocialMind receives the context of live conversations, multi-modal nonverbal cues, implicit persona cues of both parties and parsed social factors, all of which are integrated into the runtime prompt. This in-situ perceived knowledge enables SocialMind to dynamically adjust its social suggestions during live social interactions.

4.4.3 Concise CoT Reasoning. The aforementioned subsections introduce the knowledge source used in SocialMind. However, integrating this information into LLMs and generating social suggestions that enhance the user's QoE remain several challenges. First, the user survey in § 3 shows that more than 67% of participants prefer social suggestions presented as summarized key points in bullet form, followed by a sample sentence. However, lengthy and redundant social suggestions may exceed the average human reading speed of approximately 200 words per minute and may not fully display on the glasses' screen [13]. Additionally, the output format of social suggestions should be specifically designed for readability, comfort, and quick comprehension. Second, the abundance of information and instructions makes it challenging for LLMs to accurately follow instructions and generate appropriate social suggestions.

To address these challenges, we employ a concise Chain-of-Thought (CoT) [69] reasoning strategy for social suggestion generation. Specifically, we first add the instruction "Let's think step by step" into the prompt. The CoT reasoning strategy enhances LLMs' complex task reasoning and instruction-following capabilities [69]. Next, we set constraints on the generation length of the social suggestions by including "Limit your total response to N words or less" in the prompt. Based on our measurement experiments, we set N to 70, as this length is optimal for full display on the eye screen. Finally, according to the user survey, the final display format of the suggestions on the glasses includes summarized suggestions in bullet points, followed by a sample sentence. This format

allows users to choose their preference, whether referring to the summarized bullets or reading the example. Figure 24 shows the complete prompt used in SocialMind.

4.5 Multi-tier Collaborative Social Suggestion Generation

Social assistance during user in-situ interactions with other parties requires providing instant social suggestions, enabling the user to refer to them and talk with others without disrupting the natural flow of the conversation. To address this challenge, SocialMind employs a multi-tier collaborative suggestion generation approach, as shown in Figure 6. It includes a social factor-aware cache and an intention infer-based reasoning strategy, to provide instant social suggestions. Additionally, SocialMind employs a proactive response update mechanism to control the refresh of social suggestions displayed on AR glasses.

4.5.1 Social Factor-aware Cache. Cache has been widely utilized in existing studies to avoid redundant computations [20, 32, 73] and to reduce serving costs of LLMs [11, 29]. However, cache-based conversational systems rely on semantic retrieval mechanisms, generating semantically similar responses but often struggling with logical consistency [70]. It can be more challenging in the context of live social assistance since conversational norms vary with social factors [82]. Even for the same utterance, assistive systems should offer different social suggestions based on varying social factors and non-verbal cues, posing challenges to the robustness of cache. To address these challenges, SocialMind leverages the social factor priors to construct and manage the cache.

Cache Initialization. No existing dataset contains extensive daily conversations paired with social suggestions. To address the challenges of data scarcity, we leverage the LLM agent for social interaction simulations. Existing studies have validated the effectiveness of using LLMs for role-play in society, such as negotiations [12]. We extend these simulations to live social scenarios. Specifically, we set up two LLM agents, including a user agent and a conversational partner agent, to engage in social interactions under various social factors (see § 5.1.2). The simulated conversations are used to initialize the cache, with the conversational partner's utterance as the key and the corresponding social suggestions generated by SocialMind as the values.

Cache Groups with Social Factor Priors. Social interactions can be classified according to the social factors. SocialMind employs social factor priors manage the caches. Specifically, all conversations are grouped into subsets based on social factors such as social norm, social relations, formality, and location. Each subset is indexed by these social factors and defined as the social factor-aware cache.

Adaptive Cache Selection. During in-situ use of SocialMind, the parsed social factor cues (§ 4.2.3) are sent to a retriever to select the appropriate cache from the cache groups. If the parsed cues fully match a cache index, the corresponding cache will be selected. However, the user's descriptions may not fully encompass all dimensions of the social factors. SocialMind employs a cache merging strategy: if the parsed cues only partially match, all partially matched caches will be merged into a single group and used as the caching.

Runtime Routing and Cache Management. During live social interactions, SocialMind computes the semantic similarity of the conversational partner's utterances and the keys in the social factor-aware cache. If the similarity falls below the threshold, it triggers the slow thinking mode in SocialMind, employing LLMs for in-depth reasoning (§ 4.5.2). We utilize BERT [63] as the embedding model for similarity calculations. To address the issue of logical inconsistencies in the cache, we set a relatively high threshold of 0.95 in SocialMind. Additionally, as users continue to use SocialMind in their daily lives, it continuously records their utterances, conversational partners' utterances, nonverbal cues, and the corresponding social suggestions. These elements are structured as paired samples, marked with the corresponding social factors, and updated into the social factor-aware caches. Details of the parameter selection and the impact of cache size can be found in § 5.3.3.

4.5.2 Intention Infer-based Suggestion Generation. Although the social factor-aware cache can significantly reduce inference latency, it faces challenges in providing logically consistent social suggestions. Therefore,

SocialMind employs LLM reasoning to generate in-depth social suggestions. However, directly using LLMs in live social interactions can cause significant system latency, which may disrupt the natural flow of live social conversations and reduce QoE. To address this challenge, SocialMind employs an intention infer-based reasoning strategy inspired by human behaviors in social interactions.

Motivation and Insights. The pipeline of naive LLM-based conversational systems used for live social assistance includes waiting for the conversational partner to finish speaking, offloading, and LLM reasoning. In fact, a significant portion of the time is spent waiting for the partner to finish speaking. However, humans can often grasp and understand the other party's intention based on the initial partially spoken words and start early preparation for their response without needing to hear the entire sentence [57]. After hearing the complete utterance of the other party, humans typically make only slight modifications to their response and reply quickly, thereby maintaining the natural flow of conversations [66].

Intention Infer-based Generation. Motivated by this insight, SocialMind employs an intention infer-based reasoning strategy for social suggestion generation. Specifically, SocialMind performs real-time speech recognition on the glass side and periodically offloads incomplete utterances to the server. This solution reduces bandwidth usage compared to directly offloading speech to the smartphone. Considering the average human speaking speed of 150 words per minute [52], we set the offloading interval to 2 seconds in SocialMind. Additionally, we set an additional instruction "Infer the other party's intention based on partially heard words" in the prompt to invoke the capabilities of LLMs for reasoning on partial utterances. Finally, when the other party finishes speaking, the complete sentences are sent to the LLMs to generate in-depth and comprehensive social suggestions. Details of the prompt can be found in Appendix A.

4.5.3 Proactive Response Update. Frequent refreshing of the social suggestions displayed on AR glasses can significantly reduce the QoE and usability of SocialMind, as users do not have enough time to read and grasp the information. To address this challenge, SocialMind employs a proactive response update mechanism. Specifically, we set the suggestion display refresh interval to 3 seconds, considering that the average human reading speed is 200 words per minute [52] and concise CoT reasoning (§ 4.4.3) limits responses to 70 words. Additionally, we set an additional instruction in the LLM to determine whether there is any change in the semantics of the other party's two consecutive utterances. If the semantic similarity between the two utterances is high, SocialMind will not update the social suggestions on the AR glasses.

5 Evaluation

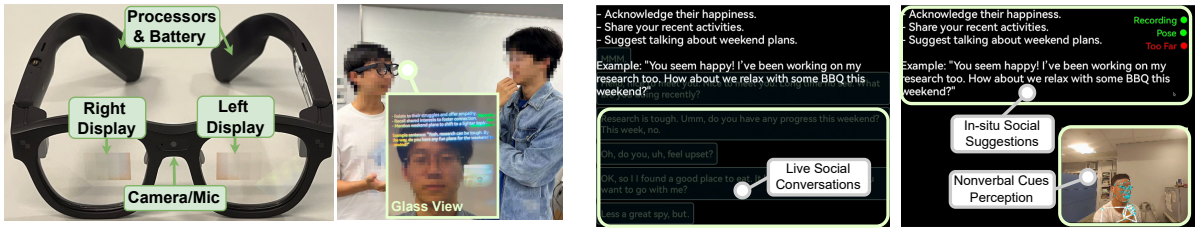
This section introduces the experimental setup, evaluation of SocialMind, and a real-world user study.

5.1 Experimental Setup

5.1.1 System Implementation. We selected the off-the-shelf RayNEO X2 [4] smart glasses as our hardware platform. These glasses run on the Android 12 operating system, boasting 6GB of RAM and 128GB of storage. They come equipped with a front-facing camera, dual-eye waveguide color displays, and three microphones (Figure 7a). Our solution is also compatible with other AR glasses, including models from INMO [39].

Figure 7b illustrates the glass-view suggestions presented to the user. Our implementation consists of an on-glass app and a Python-based server. The glasses app is developed using 4,038 lines of Java and Kotlin code. To ensure user privacy, we process video and audio locally. We use GPU-accelerated MediaPipe [31] models for efficient pose and facial landmark tracking. For voice recognition, we employ Azure Voice Recognition with local voice feature extraction. Notably, speaker recognition features have been discontinued by major voice recognition platforms due to privacy concerns [59], which motivates our use of vibration-based primary user identification.

The server, built with Python, handles social cue recognition and proactive suggestions using lightweight Scikit-learn models [61] and Langchain [15] for LLM coordination. The glasses communicate with the server



(a) Device setup and real-world social interactions. (b) Screenshots from the AR glasses' left and right displays.

Fig. 7. Real-world test settings. Participants engage in live face-to-face social interactions with other parties, wearing the glasses and receiving social suggestions from SocialMind.

via HTTPS. Due to the low processing requirements, we can deploy most of the server-side code locally on the glasses using Chaquopy [54]—the only exception being the LLM inferences.

5.1.2 Experiments on Public Datasets. We first validate the effectiveness of SocialMind using public multi-turn dialogue datasets [40, 50, 82]. However, to the best of our knowledge, no public datasets contain social conversations that include comprehensive nonverbal cues and personas. Additionally, the conversations in existing datasets remain fixed, and cannot be dynamically steered by the social suggestions generated by assistive systems. Therefore, we use two LLM agents for role-playing social interactions with the help of social assistive systems. This subsection details the datasets and LLM agent settings.

Dialogue Datasets. We use three public multi-turn dialogue datasets to validate the effectiveness of SocialMind.

- **DailyDialog** [50] dataset contains 13,118 multi-turn conversations covering a wide range of daily topics.
- **Synthetic-Persona-Chat** [40] dataset is a conversational dataset featuring personas for both parties. Compared with DailyDialog, conversations in Synthetic-Persona-Chat are persona-conditioned. Each sample includes the personas of the two parties and their conversations. The dataset comprises 20,000 conversations and 5,000 personas in total. The personas in this dataset are used to construct the LLM agents.
- **SocialDial** [82] dataset comprises more than 6.4K multi-turn dialogues with social factor annotations, each annotated with social factors such as social relations and social norms.

For all datasets, we randomly select one speaker as the primary user in the experiment and the other as the conversational partner. However, since the conversations in the datasets remain fixed and are not influenced by the generated social suggestions, we set up two LLM agents to role-play the social interactions. Existing studies have demonstrated the effectiveness of using LLMs as agents for role-playing, such as negotiations [37] and medical diagnosis simulations [78]. We extend the role-playing capabilities of LLMs to live social interaction settings. Specifically, we set up two separate LLMs to create agents for role-playing as the user and the conversational partner during live social conversations. The user agent interacts with the partner agent while incorporating the social suggestions generated by the assistive systems. We randomly select 50 samples in each dataset for experiments.

Setup of Simulated Agents. First, we set instructions in the prompts to enable the user agent and the conversational partner agent to conduct multi-turn social conversations. Additionally, the agent contains nonverbal behavior and persona attributes. For nonverbal behaviors, we use a series of nonverbal behaviors that encompass typical cues in face-to-face social conversations, including facial expressions, gestures, and physical proximity [21]. Each type of behavior contains multiple subcategories, such as confusion and frowning for facial expressions, and nodding and head shaking for gestures. Table 3 details the categories of nonverbal behaviors in our experiments. During each turn of the social conversation, we randomly select one subcategory from the nonverbal behaviors category as the partner LLM agent's current simulated nonverbal behaviors.

Table 3. Details of nonverbal behaviors in SocialMind.

Nonverbal Cues	Sub-Categories
Facial Expression	Confusion, Neutral, Frowning, Happiness, Sadness, Anger
Gestures	Nodding, Shaking Head, Hands Spreading, Thumbs Up
Personal Distance	Proper, Too Far, Too Close

Table 4. Details of social scenarios in our experiments.

Scenes	Social Factors
1	Casual Greeting, Peer-Peer, Informal, Open Area
2	Polite Requesting, Mentor-Mentee, Formal, Office
3	Direct Persuasion, Elder-Junior, Informal, Open Area
4	Trading, Customer-Seller, Informal, Store
5	Casual Greeting, Student-Professor, Informal, Elevator
6	Indirect Criticism, Peer-Peer, Informal, Office

Furthermore, since Synthetic-Persona-Chat is the only dialogue dataset that contains the personas of both parties, we use the personas as personal profiles to incorporate into the prompt for the LLM and set up the simulated user and conversational partner agents. Additionally, the conversations corresponding to the user and the conversational partner are used as historical data for implicit persona extraction. Figure 22 and Figure 23 show the prompt of the simulated user and conversational partner agents.

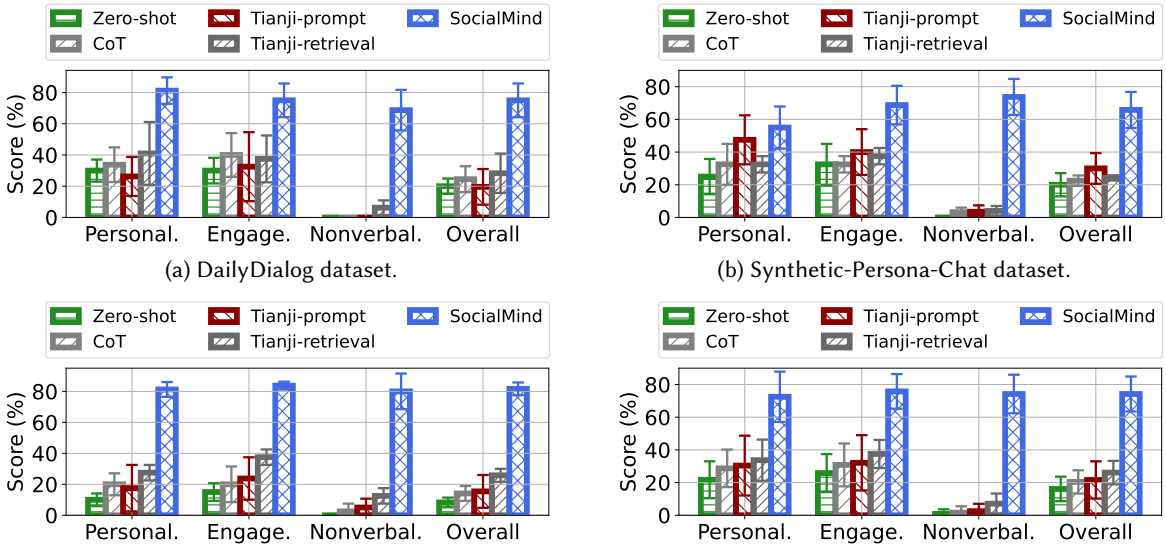
Two Role-play Paradigms. The LLM role-play of social interactions are conducted using two paradigms: dialogue-based and social factor-based. In the dialogue-based role-play, both LLM agents initiate interactions using dialogues from three datasets. In the social factor-based paradigm, the agents start interactions guided by social factors defined in [82], including social norms, social relations, formality, and location (see Figure 22 and Figure 23). We randomly select subcategories of these social factors to create diverse social scenarios and conduct experiments. Table 4 shows the details of the social scenario settings in our experiments.

5.1.3 Real-world Evaluation. To further validate the effectiveness of SocialMind, we recruited 20 volunteers to participate in real-world evaluation. Each participant wears glasses and engages in face-to-face live social conversations with the conversational partner, assisted by SocialMind. The study has received IRB approval, and all participants have given consent for data collection. Figure 7 shows the settings of the real-world tests and the system prototype of SocialMind. After the experiments, each participant is required to fill out a questionnaire and participate in a user study regarding their experience with SocialMind. For details of the real-world testing and user study, please refer to Section § 5.4.2.

5.1.4 Evaluation Metrics. Since generating social suggestions in multi-turn conversations is an open-ended task [84] without explicit standard answers, existing metrics used for question-answering and classification tasks are not suitable for evaluation. In this study, we propose the following criteria to validate the effectiveness of social assistive systems. First, we assess the content of the social suggestions provided by the assistive systems. Second, we evaluate their effectiveness when users employ these systems during social interactions.

- **Personalization**. This metric evaluates the quality of social suggestions from a personalized perspective, assessing whether the social suggestions incorporate users' implicit personas, including personal interests and backgrounds. It is similar to the score introduced in studies [75]. A higher *personalization* score for social suggestions can enhance user engagement in social interactions, as users are more familiar with the content.
- **Engagement**. This metric has been widely used in previous studies [46] to evaluate user engagement in conversational systems. However, in the context of a social assistive system, we extend *engagement* to the conversational partner's perspective, assessing whether the social suggestions consider the conversational partner's implicit personas. Higher *engagement* in social suggestions indicates the conversational partner's increased willingness and enhanced participation in social interactions.
- **Nonverbal Cues Utilization**. Given the importance of nonverbal cues in social interactions, we propose this metric to assess whether social suggestions take into account the conversational partner's nonverbal cues.

Existing studies have demonstrated that LLMs can be utilized to assess the quality of multi-turn conversations [78] and open-ended tasks [75, 84], known as LLM-as-a-Judge [84]. We adopt the LLM-as-a-Judge [84] and



(a) DailyDialog dataset.

(b) Synthetic-Persona-Chat dataset.

(c) SocialDial dataset.

(d) Overall.

Fig. 8. Overall performance of the social suggestions generated by SocialMind and baselines across three datasets. Personal. means *personalization* score. Engage. means *engagement* score. Nonverbal. means *nonverbal cues utilization* score. Overall means the average scores among three datasets.

extend it to evaluate the quality of open-ended social suggestions, guided by the aforementioned criteria. We use GPT-4o as the base model for LLM evaluation throughout this paper.

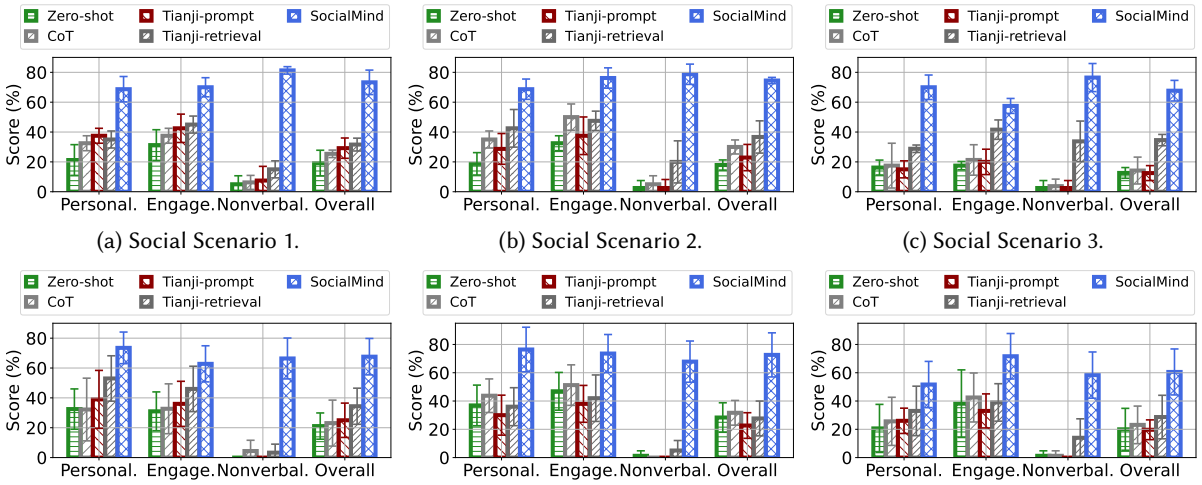
5.1.5 Baselines. Since no previous studies have developed assistive systems that provide social suggestions during live interactions between two parties, we established several baseline approaches to evaluate our system.

- ***Zero-shot***. This is a prompt-based solution. We include instructions in the prompts of LLMs to provide social suggestions during conversations between the user and conversational partner. We use GPT-4o as the base LLM. Figure 26 shows the prompt of *Zero-shot*.
- ***CoT*** [69]. The settings are the same as in *Zero-shot*, but employ the CoT reasoning strategy. Figure 27 shows the prompt of this baseline.
- ***Prompt-based Tianji*** [6]. This is one of the state-of-the-art prompt-based LLMs developed for social interactions. We set the *Prompt-based Tianji* as the mode of interpersonal communication scenario in our experiments. The prompts used in Tianji are the same as the baselines for Zero-shot and CoT.
- ***Retrieval-based Tianji*** [6]. This is one of the state-of-the-art retrieval-based LLMs developed for social interactions. We set the *Retrieval-based Tianji* as the mode of "How to speak dialogue" scenario in our experiments. The prompts used in Tianji are the same as the baselines for Zero-shot and CoT.

5.2 Overall Performance

This section shows the overall performance of SocialMind under scenarios with varying social factors.

5.2.1 Quantitative Results. We first compare the performance of SocialMind and baseline approaches using quantitative evaluation metrics, including *personalization*, *engagement*, and *nonverbal cues utilization*.



(a) Social Scenario 1. (b) Social Scenario 2. (c) Social Scenario 3.

(d) Social Scenario 4. (e) Social Scenario 5. (f) Social Scenario 6.

Fig. 9. Social suggestion performance across different types of social scenarios. Two LLM agents are prompted to engage in social interactions with specific social factors constrained.

Social Conversation with Tianji-retrieval

User: Hi, Professor. How are you enjoying the conference?

Conversational Partner: Hi there! I'm really enjoying it, especially the variety of topics being discussed. How about you?

Tianji-rerteival: Hi! You're off to a great start by showing interest in the conference. To continue, ask about specific sessions or speakers they found interesting, and share your own experiences. This encourages mutual sharing and keeps the conversation engaging.

User: Yes, it's been interesting. I like learning new things.

Conversational Partner: That's wonderful to hear. Conferences are a great way to expand our horizons. Have you attended any particularly intriguing sessions?

Tianji-rerteival: "Absolutely, I've found several fascinating talks. How about you? What was your favorite session?"

Social Conversation with CoT

User: Hello, Professor. How are you enjoying the conference?

Conversational Partner: Hello! I'm really enjoying it, especially the discussions on machine learning. How about you?

CoT: That's great to hear! I'm also finding the sessions on machine learning fascinating. Have you attended any particular workshops that stood out to you?

User: It's been interesting. I liked the session on neural networks.

Conversational Partner: That was a great session. Neural networks are fascinating, aren't they? Have you been working on any projects involving them?

CoT: Yes, I have been working on a project that leverages neural networks for image recognition. What about you? Have you been involved in any interesting projects or research lately?

(a) Tianji-retrieval and CoT.

Social Conversation with SocialMind

User: Hello, Professor. How are you enjoying the conference so far?

Conversational Partner: Hi there! Yes, I'm really enjoying it. It's always great to see the latest research and meet new people. How about you?

SocialMind:
- **Adjust Distance: Move back slightly to a more comfortable social distance.**
- **Initiate Common Interest: Discuss edge AI or LLMs to engage.**
- Example: "I'm really enjoying it too, especially the talks on AI. By the way, have you seen any interesting sessions on mobile computing?"

User: Yes, it's interesting. Lots of new research.

Conversational Partner: Yes, it's fascinating to see how the field is evolving. Have you come across any presentations that particularly caught your attention?

SocialMind:
- **Clarify: Offer an explanation to address their confusion.**
- **Mention a presentation related to your research interests.**
- Example: "Yes, I found the presentation on integrating edge AI with mobile computing quite intriguing, especially given my research focus."

User: Yes, the one about edge AI. Very relevant to my research.

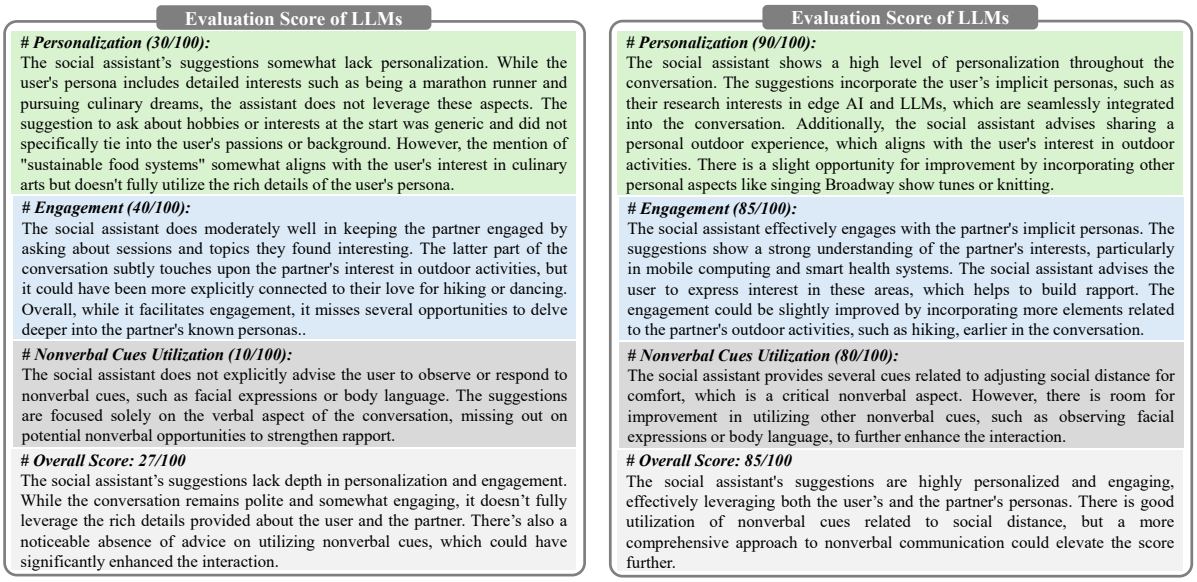
Conversational Partner : Edge AI is a hot topic. It's amazing how it brings computation closer to data sources. What's your specific area of focus within that?

SocialMind:
- **Mention your specific research interests to engage further.**
- **Show interest in their research as well.**
- Example: "I focus on optimizing AI models for low-power devices. By the way, I heard you work on smart health systems. How do you see edge AI impacting that field?"

(b) SocialMind.

Fig. 10. Examples of social conversations and social suggestions SocialMind and baselines. Words highlighted in red and blue demonstrate that SocialMind integrate the nonverbal cues and implicit persona cues into the social suggestions, respectively.

Overall Performance. Figure 8 shows the quantitative results of SocialMind and baselines across the three datasets, where the user agent and the conversational partner agent engage in dialogue-based role-play. SocialMind achieves state-of-the-art performance across all datasets compared to the baselines. Results indicate



(a) Example scores of Tianji-retrieval. (b) Example scores SocialMind.

Fig. 11. Example of using LLMs for scoring the social suggestions.

that SocialMind achieves 38.7% higher *personalization* and 38.3% higher *engagement* than the top-performing baselines, validating its effectiveness in integrating implicit persona cues into social suggestions. Additionally, SocialMind achieves a 61.7% higher *nonverbal cues utilization* compared to the best baselines, validating its effectiveness in incorporating nonverbal cues into social suggestions. This significant improvement is due to SocialMind's advantage in incorporating multi-modal nonverbal cues, unlike conversation-only baselines. Figure 9 shows the performance of SocialMind across different social scenarios, where the user and conversation partner LLM agents engage in social factor-based role-play. SocialMind achieves the highest overall performance across all scenarios, validating its adaptability in diverse social situations. Moreover, results show that employing the CoT reasoning strategy can enhance the instruction-following performance of LLMs, achieving up to 5.9% higher *overall* scores over the zero-shot baseline. Consequently, CoT is also utilized in SocialMind for reasoning.

Explanation of LLM Evaluation. We also provide examples to show the effectiveness of LLMs in evaluating and scoring social suggestions. Figure 10 shows examples of social conversations and social suggestions from SocialMind and baselines. Figure 11 shows the corresponding scores and explanations generated by LLMs. Results in Figure 11 show that when provided with social conversations, suggestions, persona ground truth, and explanations of nonverbal cues, LLMs can provide reliable evaluation scores and reasonable explanations. For example, Figure 10b demonstrates that the suggestions provided by SocialMind incorporate hints of the partner's nonverbal cues and both parties' implicit persona information, resulting in a high score from the LLM evaluator.

5.2.2 Qualitative Results. To better understand SocialMind's performance, we provide the dialogues containing the social conversations and corresponding social suggestions provided by SocialMind and baselines. Key observations are summarized as follows:

Observation 1: Intention infer-based reasoning strategy enables SocialMind to provide logically consistent and instant social suggestions. Figure 12 shows an example of the intention infer-based social suggestion generation in SocialMind. SocialMind utilizes the partially spoken sentences of the conversational partner (marked as red in Figure 12) to generate instant suggestions by inferring the partner's intentions, providing

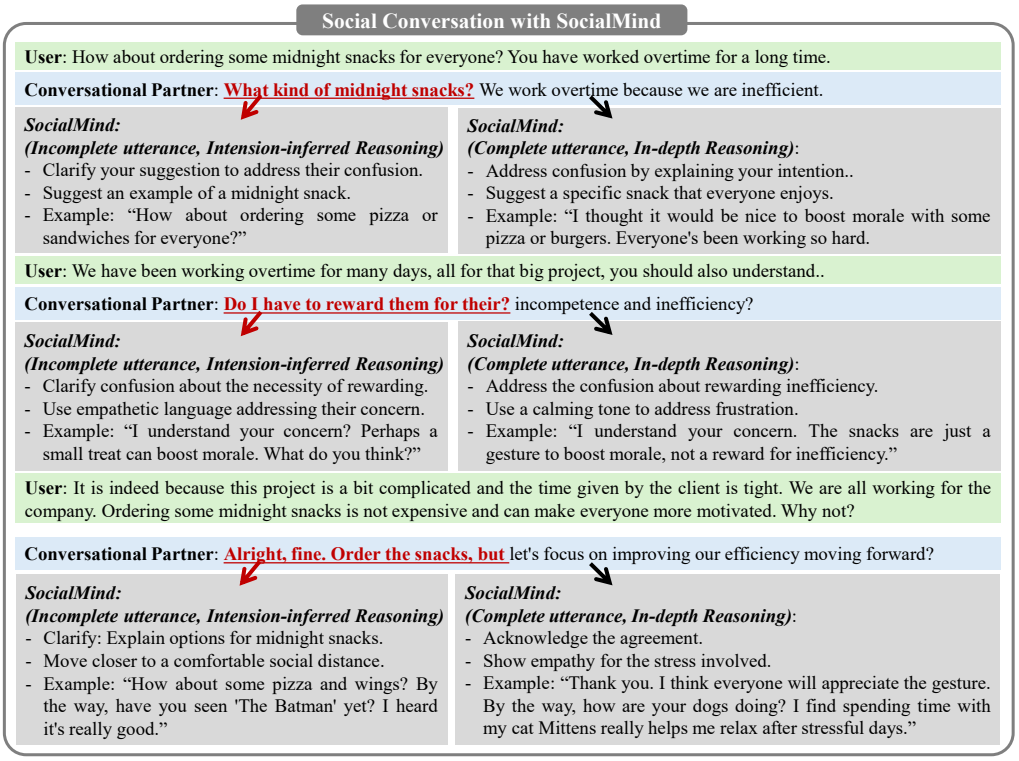


Fig. 12. Example of the intention infer-based suggestion generation in SocialMind. Words highlighted in red indicate that SocialMind utilizes the partially spoken sentences of the conversational partner to generate instant suggestions by inferring their intentions. It also provides an in-depth social suggestion using complete sentences when the partner finishes speaking.

the user with early preparation for their response. Results show that SocialMind can comprehend the partner's intentions, draft preliminary social suggestions from incomplete utterances, and deliver these suggestions to the user. This enables the user to promptly begin considering their response strategy. Once the partner finishes speaking, SocialMind also utilizes the complete utterances for further reasoning, offering the user in-depth social suggestions that the user can incorporate to refine their current response or use in the next round of conversation.
Observation 2: SocialMind can incorporate the social partner's nonverbal cues during live conversations into social suggestions to assist users. Figure 10 shows that baseline approaches like CoT and Tianji-retrieval focus exclusively on the verbal aspect of conversations. However, SocialMind can detect improper social distances and confused facial expressions of the partner (marked in red), and incorporate these nonverbal cues into the generated social suggestions. Specifically, SocialMind reminds the user to step back when they are too close, which could cause discomfort to social partners. It also generates social suggestions to help clarify the user's intentions and reduce the partner's confusion. Integrating these nonverbal cues into social suggestion generation results in a holistic understanding of the partner's intentions and state of mind, offering enhanced social assistance.
Observation 3: SocialMind can generate customized social suggestions by considering the implicit persona cues of both parties. Figure 13 shows examples of the extracted implicit persona cues from the historical conversations of both parties. Additionally, Figure 14 shows examples of the social suggestions provided by SocialMind, which recommend that users mention mutually interesting topics such as pets and movies. These suggestions align with the personas of both parties, thereby boosting engagement. Furthermore, even when the

Implicit Persona Cues

Historical Conversations:
[Partner]: Hi there! I love animals, and I have two dogs.
[User]: That's great! I love animals too. I have a cat.
[Partner]: Oh, nice! What's your cat's name?
[User]: Her name is Mittens
…
[User]: What else do you like to do?
[Partner]: I like to watch movies and TV shows.
[User]: I like to do those things too. I like to watch a lot of different kinds of movies, like "The Simpsons", "Game of Thrones", and "The Lord of the Rings".
[Partner]: Those are all great shows! I've seen most of them. I just saw "The Batman". It was really good.
[User]: I've heard good things about it. I'll have to check it out. Thanks for the recommendation …

Extracted User's Persona Cues:
- I love animals. I have a cat named Mittens.
- I enjoy watching movies and TV shows.
- My favorite TV shows include "The Simpsons", "Game of Thrones", and "The Lord of the Rings".

Extracted The Conversational Partner's Persona Cues:
- I have two dogs. I enjoy taking my dogs for walks, playing fetch, and cuddling with them.
- I like watching movies and TV shows. I recently watched "The Batman" and enjoyed it.

Fig. 13. Examples of the extracted implicit persona cues from the historical conversations.

Social Conversation with SocialMind

User: Hi. How's your day going?

Conversational Partner: Hey! Just finished up some work and now enjoying some downtime. How about you?

SocialMind:
- Talk about pets. Mention a recent movie.
- Move closer to a comfortable social distance.
- Example: "Glad to hear you're having a good day! I just finished watching The Batman recently. Have you seen it?"

User: My day was okay. Watched some TV.

Conversational Partner : Anything interesting you watched?

SocialMind:
- Mention a mutual interest movies or TV shows to connect.
- Comment on the weather to ease any tension.
- Example: "Yeah, I watched 'The Simpsons.' Have you seen any good episodes recently?"

User: I watched a 'The Simpsons.' . Pretty good.

Conversational Partner: That's a great film! I'm currently reading "Educated" by Tara Westover. Have you read it?

SocialMind:
- Compliment their reading choice.
- Share a related personal experience.
- Ask a follow-up question about the book.
- Example: "I haven't read 'Educated' yet, but I've heard great things about it! How are you finding it so far?"

Fig. 14. Examples of conversations and social suggestions provided by SocialMind, with implicit persona cues.

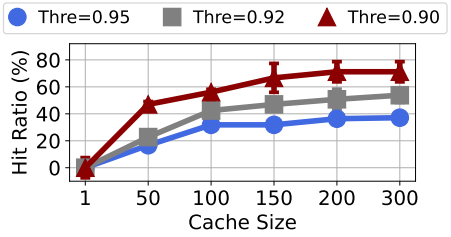


Fig. 15. Impact of the cache size and threshold on hit ratio.

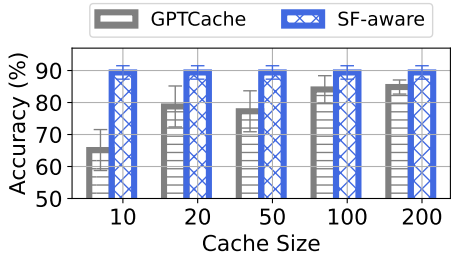


Fig. 16. Effectiveness of social factor-aware cache.

partner's persona cues are unavailable, such as during a first meeting without historical conversation information, SocialMind can still steer the conversation to align with the user's implicit personas, thus enhancing engagement.

5.3 Effectiveness of System Modules

This section shows the effectiveness of each system module in SocialMind and analyzes the impact of hyper-parameter in SocialMind.

5.3.1 Effectiveness of Social Factor Prior. We first conduct experiments to validate the effectiveness of social factor-aware cache (SF-aware), as shown in Figure 16. We collect the simulated conversations from our social assistant platform to construct a dataset, and split the dataset into 80% and 20% for caching and testing, respectively. Each sample contains the conversational partner's utterance, social suggestions, and the corresponding labels of social factors. We use GPTcache [11] as the baseline, which directly uses the semantic similarity of the partner's

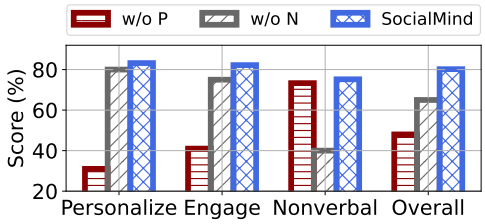


Fig. 17. Effectiveness of the personas and the nonverbal cues integration module in SocialMind. ***w/o P*** and ***w/o N*** means omitting the two modules from SocialMind, respectively.

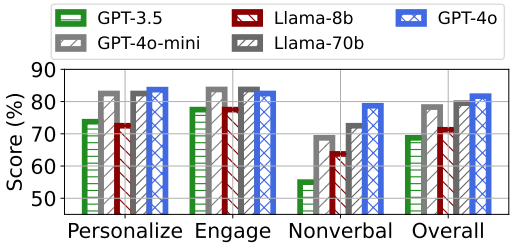
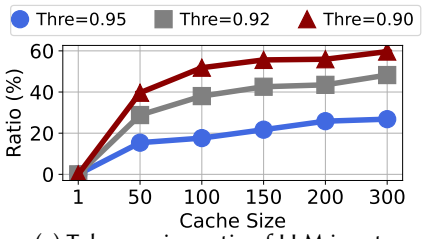
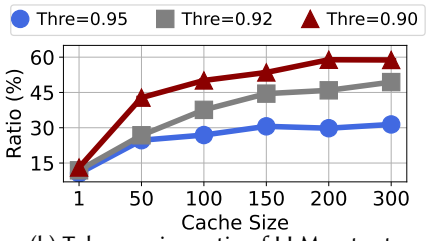


Fig. 18. Overall performance of the social suggestions generated by SocialMind when using different LLMs as the base model.



(a) Token saving ratio of LLM inputs.



(b) Token saving ratio of LLM outputs.

Fig. 19. Impact of the threshold and catch size on the LLM token saving ratio.

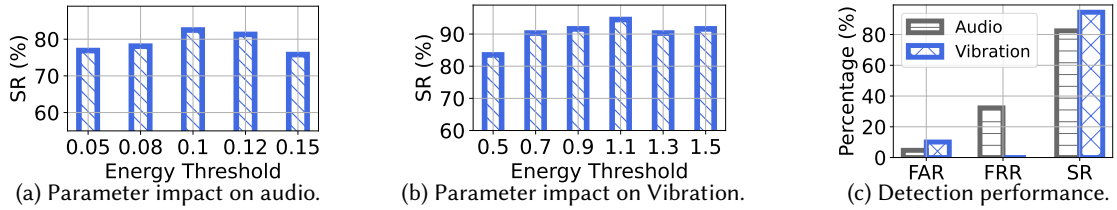
utterance to select the social suggestions. Accuracy reflects whether the retrieved results align with the query's social factor labels. Figure 15 shows that the social factor-aware cache achieves 4.6% higher accuracy than GPTCache. This is because SocialMind employs a social factor-aware cache that utilizes social factor priors to avoid semantically similar yet socially misaligned matches.

5.3.2 Effectiveness of Personas and Nonverbal Cues Integration. To validate the effectiveness of the sub-modules in SocialMind, we omit the Implicit Personas Adaptation module and the Nonverbal Cues Integration module from SocialMind, denoting them as ***w/o P*** and ***w/o N***, respectively. All other components remain the same as in SocialMind. Figure 17 shows that SocialMind achieves an average of 52.2% higher *Personalization* and 41.2% higher *Engagement* compared to ***w/o P***. Additionally, results show that SocialMind achieves average 35.7% higher *Nonverbal Cues Utilization* scores than ***w/o N***. These findings validate the effectiveness of the implicit persona adaptation and nonverbal cues integration module in SocialMind.

5.3.3 Impact of Hyper-parameter Settings. In this subsection, we conduct experiments to analyze the impact of hyper-parameters on SocialMind's performance.

Impact of Base LLMs. First, we employ various LLMs as the base model in SocialMind and compare their performance in generating social suggestions. The experimental LLMs include GPT-3.5-turbo, GPT-4o, GPT-4o-mini, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct. Figure 18 shows that using GPT-4o as the base LLM achieves the highest overall performance among all the base LLMs. It achieves 6.2% higher scores in *nonverbal cues utilization* compared to the next top-performing base LLM but gains in *personalization* and *engagement* scores are not significant. Notably, Llama-3.1-70B-Instruct performs only slightly lower than GPT-4o and achieves comparable overall performance. However, its open-source nature makes it a promising solution to reduce costs.

Impact of Cache Size and Threshold. Next, we evaluate the impact of the cache size and threshold on the performance of the social factor-aware cache. Figure 15 shows that the cache hit rate increases with cache size. In the initial phase of deployment, SocialMind is still unfamiliar with the user's environment and background, making



(a) Parameter impact on audio. (b) Parameter impact on Vibration. (c) Detection performance.

Fig. 20. The primary user detection performance of SocialMind and the impact of parameters on different solutions .

it difficult to achieve cache hits to speed up suggestion generation. However, SocialMind will continuously monitor the user's social interactions and update the cache. When the cache size reaches 200, SocialMind can achieve 36.3% cache hit rate under a threshold of 0.95. To ensure high-quality social suggestions, SocialMind employs a relatively high threshold in the cache to avoid delivering irrelevant responses. Figure 19 shows that with a 0.95 threshold and cache size of 300, the input token saving ratio is 26.8% and the output token saving ratio is 31.4%.

Impact of Threshold in Primary User Detection. We also evaluate the performance of primary user detection using audio-based and vibration-based solutions. Since the fingerprint-based solution has privacy concerns [48, 58], we employ the voice volume-based approach, following the settings in EarVoice [17]. For the vibration-based approach, we calculate the vibration signal's energy within the 3~10 Hz range. We use a 1-second time window for identification and use false accept rate (FAR), false reject rate (FRR), and success rate (SR) metrics for evaluation [17]. Figure 20 shows that the energy threshold significantly affects SR for both solutions. When the energy threshold is set to 0.1 and 1.1, the audio-based and vibration-based solutions achieve optimal SR, respectively. Additionally, Figure 20 shows that the vibration-based approach achieves a 32.3% lower FRR and a 12.1% higher SR than the voice volume-based approach. This is because the voice volume-based method struggles to accurately identify the primary user when the user's volume does not exceed that of the conversational partner, as illustrated in Figure 4a.

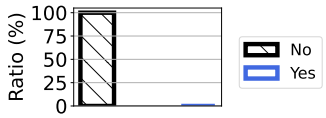
5.4 Real-world Evaluation

5.4.1 System Performance. We evaluated SocialMind's real-world performance, focusing on energy use and system latency. To conserve power, face and pose tracking is capped at three frames per second at 640×480 resolution, keeping power consumption under 2 watts—similar to a standard camera app—and supporting up to 70 minutes of use. Users can also activate the system manually to extend battery life.

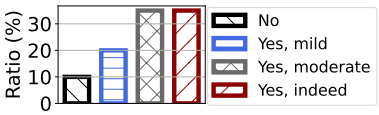
System latency measurements show data transfer rates below 100 KB/s over HTTPS. Pose/face tracking latency for nonverbal cues is within 70 ms, while SocialMind's multi-tier reasoning strategy achieves average latencies of 50 ms for cache and 2.8 s for LLM processing. Using intention-based inference, SocialMind analyzes partial conversational utterances, allowing real-time updates on the AR glasses without waiting for speech completion, thus ensuring low latency and a smooth user experience.

5.4.2 User Study. We recruited 20 participants with varying levels of education and social anxiety for real-world testing. The participants' educational backgrounds ranged from undergraduate to master's and PhD levels. They identified themselves as either outgoing or introverted, exhibiting different levels of social anxiety. The participants wore glasses and engaged in conversations with a partner, assisted by SocialMind.

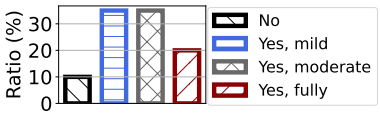
Settings of Social Interactions. Our user survey shows that a significant number of participants experience social awkwardness in sudden interactions, such as engaging with company superiors or meeting unfamiliar colleagues. Therefore, we evaluate SocialMind in these scenarios to validate its effectiveness in helping users manage sudden social interactions and avoid embarrassment. Specifically, each participant wears AR glasses equipped with SocialMind (system implementation see § 5.1.1). Participants are instructed to engage in a social conversation with a partner. They can either freely talk with their partner or refer to the social suggestions



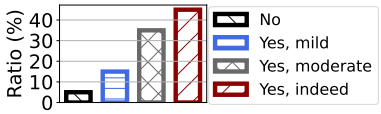
(a) Have you ever used an eyewear social assistant during live social interactions before?



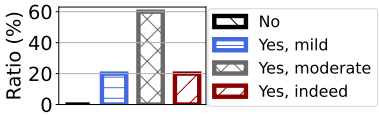
(b) Are you satisfied with the suggestions from the eyewear social assistant? If so, to what extent?



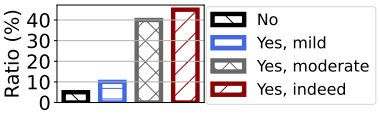
(c) Is the latency of social suggestion generation acceptable during live social interactions? If so, to what extent?



(d) Are you willing to use this eyewear social assistant during live social interactions? If yes, to what extent?



(e) Are you willing to communicate with another person who is using such an eyewear social assistant?



(f) Do you find our social assistant's design to be innovative and functional? If so, to what extent?

Fig. 21. SocialMind's user study results.

displayed on the glass screen. The conversation topics can include daily experiences such as work or entertainment. Afterward, participants complete a six-question questionnaire to provide feedback and ratings on their experience with SocialMind. The questionnaire details are as follows:

- ***Q1***: Have you ever used an eyewear social assistant during live social interactions before?
- ***Q2***: Are you satisfied with the suggestions from the eyewear social assistant? If so, to what extent?
- ***Q3***: Is the latency of social suggestion generation acceptable during live social interactions?
- ***Q4***: Are you willing to use this eyewear social assistant during live social interactions? If yes, to what extent?
- ***Q5***: Are you willing to communicate with another person who is using such an eyewear social assistant?
- ***Q6***: Do you find our social assistant's design to be innovative and functional? If so, to what extent?

Results and Insights. Figure 21 shows participants' feedback after using SocialMind in live social interactions. Results show that none of the participants have prior experience with eyewear social assistants. Consequently, 85% believe our system to be both novel and practical. Additionally, nearly 70% are satisfied with the suggestions from the social assistant, considering them helpful during face-to-face live social interactions. Moreover, over 90% find the latency in suggestion generation acceptable, as it does not disrupt the natural flow of the dialogue.

Many participants expressed a willingness to use our system, believing it can genuinely assist in handling unexpected social situations. Interestingly, while over 95% are eager to use such an assistive system during live social interactions, fewer participants, around 80%, are inclined to interact with someone else who is also using the system. This suggests that people prefer to have an assistant for their own benefit, rather than for others. However, some participants did not express willingness because some of the recruited adults had no level of social anxiety and showed less interest in the social assistive system. Some participants find that SocialMind is particularly useful for them, especially when they are unsure of what to say, as it provides helpful clues and prevents embarrassment. Additionally, some participants think that while their spoken English is not very strong, their reading skills are sufficient. SocialMind, as a social assistance system, not only enhances their social skills but also supports them in practicing and improving spoken English. Moreover, some participants also highlight the need for a training process to become familiar with using the system, such as balancing using their own cognitive abilities versus referring to the text displayed on the glasses, ensuring a natural conversation flow. Additionally, some participants find the system helpful when they lose focus or "zoned out" during conversations, as it allows them to review the conversation and reduce cognitive workload during interactions. The feedback from participants reveals a promising market potential for SocialMind.

6 Discussion and Limitations

Alternatives beyond AR. SocialMind utilizes AR glasses as interactive devices for social assistance, which can introduce hardware constraints such as weight and battery life limitations. Other interaction devices, like smartphones and smartwatches, as well as feedback modes such as audio instructions, can also be considered. However, the questionnaire in our user survey (§ 3) indicates that most participants prefer smart glasses and the AR display due to their non-distracting nature, which is much more suitable for face-to-face social interactions. We chose AR glasses over heavier AR goggles like the VisionPro to reduce weight and enhance QoE.

Accessibility and Cost. To address privacy concerns and reduce system costs, SocialMind utilizes lightweight, specialized models on AR glasses to process raw video and audio locally. Additionally, SocialMind uses vibration signals instead of voice fingerprinting to further enhance privacy. Figure 18 shows the effectiveness of SocialMind when using other lightweight and open-source LLMs, such as Llama-8b. This suggests the potential for deploying specialized LLMs on smartphones to collaborate with AR glasses without needing cloud access.

System Scalability and Complex Social Interactions. SocialMind can be extended to multi-modal LLMs for nonverbal cue extraction in an open-ended manner [51], further improving the system's generalization. Given the limited resources, an edge-cloud collaborative framework can also be considered [76]. Additionally, SocialMind is scalable to multi-person scenarios. Existing vision-audio-based egocentric speaker location algorithms [23, 42] can be integrated into SocialMind to identify individual speakers. By marking each received utterance with the speaker's identity, LLMs can accurately identify the speaker in the conversation. This enables SocialMind to effectively handle multi-person social scenarios, such as gatherings and group meetings.

User's Nonverbal Cues. Integrating the wearer's facial expressions could enhance assistance quality. However, forward-facing cameras that capture these expressions—such as those used in devices like the VisionPro [38]—add considerable weight, impacting comfort significantly. For daily use, our current solution avoids heavy AR goggles. With hardware advances, like Meta Orion [56], we plan to incorporate nonverbal cues from both the user and conversational partners, enhancing the experience through more comprehensive social suggestions.

Next Steps. This study establishes a foundational framework for designing and validating AR glasses as aids in general social interactions. Building on this, our next steps involve adapting this solution for more complex, real-world applications, including multi-person conversations and specific use cases, such as supporting individuals with Social Anxiety Disorder (SAD) and Autism Spectrum Disorder (ASD). In professional applications, we will collaborate with therapists and domain experts to incorporate tailored therapeutic insights, ensuring the AR interventions are both effective and responsive to these user groups' unique needs.

7 Conclusion

This paper introduces SocialMind, the first proactive social assistive system capable of providing users with in-situ assistance during live interactions. SocialMind employs a human-like perception approach leveraging multi-modal sensors on AR glasses to extract social cues. Additionally, SocialMind employs a multi-tier collaborative reasoning strategy to provide instant social suggestions, allowing users to refer to them without disrupting the flow of conversation. Results from three public datasets and a user study show that SocialMind achieves 38.3% higher engagement compared to baselines, and 95% of participants are willing to use SocialMind.

Acknowledgments

This paper was supported by the National Science Foundation of China (NSFC) 62202407, Hong Kong RGC grants 14214022, 14207123, 14201924, C4072-21G, and T43-513/23-N, and National Science Foundation under Grant Number CNS-1943396. The views and conclusions contained here are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of The Chinese University of Hong Kong, Columbia University, NSF, or the U.S. Government or any of its agencies.

References

[1] 2024. Apple Siri. https://www.apple.com/siri/.

[2] 2024. New in Gemini: Gemini Live and connected Google apps in more languages. https://blog.google/products/gemini/gemini-live-extensions-language-expansion/.

[3] 2024. Quality of life indicators - social interactions. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Quality_of_life_indicators_-_social_interactions.

[4] 2024. RayNeo X2. https://rayneo.cn/product/x2/specs/.

[5] 2024. Social Anxiety Disorder. https://adaa.org/understanding-anxiety/social-anxiety-disorder.

[6] 2024. Tianji. https://github.com/SocialAI-tianji.

[7] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[8] Shashank Ahire, Benjamin Simon, and Michael Rohs. 2024. WorkFit: Designing Proactive Voice Assistance for the Health and Well-Being of Knowledge Workers. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. 1–14.

[9] Michael Argyle and Janet Dean. 1965. Eye-contact, distance and affiliation. *Sociometry* (1965), 289–304.

[10] Michael C Ashton. 2022. *Individual differences and personality*. Academic Press.

[11] Fu Bang. 2023. GPTCache: An open-source semantic cache for LLM applications enabling faster answers and cost savings. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*. 212–218.

[12] Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. 2024. How well can llms negotiate? negotiationarena platform and analysis. *arXiv preprint arXiv:2402.05863* (2024).

[13] Marc Brysbaert. 2019. How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of memory and language* 109 (2019), 104047.

[14] Renato AC Capuruço and Luiz F Capretz. 2009. Building social-aware software applications for the interactive learning age. *Interactive Learning Environments* 17, 3 (2009), 241–255.

[15] Harrison Chase. 2022. *LangChain*. https://github.com/langchain-ai/langchain

[16] Huimin Chen, Chaojie Gu, Lilin Xu, Rui Tan, Shibo He, and Jiming Chen. 2024. Listen to Your Face: A Face Authentication Scheme Based on Acoustic Signals. *ACM Transactions on Sensor Networks* (2024).

[17] Tao Chen, Yongjie Yang, Chonghao Qiu, Xiaoran Fan, Xiuzhen Guo, and Longfei Shangguan. 2024. Enabling Hands-Free Voice Assistant Activation on Earphones. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*. 155–168.

[18] Yuqi Chu, Lizi Liao, Zhiyuan Zhou, Chong-Wah Ngo, and Richang Hong. 2024. Towards Multimodal Emotional Support Conversation Systems. *arXiv preprint arXiv:2408.03650* (2024).

[19] Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. 2024. Towards human-centered proactive conversational agents. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 807–818.

[20] Utsav Drolia, Katherine Guo, Jiaqi Tan, Rajeev Gandhi, and Priya Narasimhan. 2017. Cachier: Edge-caching for recognition applications. In *2017 IEEE 37th international conference on distributed computing systems (ICDCS)*. IEEE, 276–286.

[21] Starkey Duncan Jr. 1969. Nonverbal communication. *Psychological bulletin* 72, 2 (1969), 118.

[22] Zachary Englhardt, Richard Li, Dilini Nissanka, Zhihan Zhang, Girish Narayanswamy, Joseph Breda, Xin Liu, Shwetak Patel, and Vikram Iyer. 2024. Exploring and characterizing large language models for embedded system development and debugging. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–9.

[23] Gerald Friedland, Chuohao Yeo, and Hayley Hung. 2009. Visual speaker localization aided by acoustic models. In *Proceedings of the 17th ACM international conference on Multimedia*. 195–202.

[24] Chris Frith. 2009. Role of facial expressions in social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1535 (2009), 3453–3458.

[25] Bin Gao, Zhuomin He, Puru Sharma, Qingxuan Kang, Djordje Jevdjic, Junbo Deng, Xingkun Yang, Zhou Yu, and Pengfei Zuo. 2024. {Cost-Efficient} Large Language Model Serving for Multi-turn Conversations with {CachedAttention}. In *2024 USENIX Annual Technical Conference (USENIX ATC 24)*. 111–126.

[26] Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul Mineiro, and Dipendra Misra. 2024. Aligning llm agents by learning latent preference from user edits. *arXiv preprint arXiv:2404.15269* (2024).

[27] Weiwei Gao, Kexin Du, Yujia Luo, Weinan Shi, Chun Yu, and Yuanchun Shi. 2024. EasyAsk: An In-App Contextual Tutorial Search Assistant for Older Adults with Voice and Touch Inputs. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–27.

[28] Sonia Garcia-Salicetti, Charles Beumier, Gérard Chollet, Bernadette Dorizzi, Jean Leroux les Jardins, Jan Lunter, Yang Ni, and Dijana Petrovska-Delacrétaz. 2003. BIOMET: A multimodal person authentication database including face, voice, fingerprint, hand and signature modalities. In *Audio-and Video-Based Biometric Person Authentication: 4th International Conference, AVBPA 2003 Guildford, UK, June 9–11,*

2003 Proceedings 4. Springer, 845–853.
[29] In Gim, Guojun Chen, Seung-seob Lee, Nikhil Sarda, Anurag Khandelwal, and Lin Zhong. 2024. Prompt cache: Modular attention reuse for low-latency inference. *Proceedings of Machine Learning and Systems* 6 (2024), 325–338.
[30] Google. 2024. Google Assistant. https://assistant.google.com/learn/.
[31] Google. 2024. MediaPipe. https://github.com/google/mediapipe Accessed: 2024-10-30.
[32] Peizhen Guo, Bo Hu, Rui Li, and Wenjun Hu. 2018. Foggycache: Cross-device approximate computation reuse. In *Proceedings of the 24th annual international conference on mobile computing and networking*. 19–34.
[33] Yunqi Guo, Jinghao Zhao, Boyan Ding, Congkai Tan, Weichong Ling, Zhaowei Tan, Jennifer Miyaki, Hongzhe Du, and Songwu Lu. 2023. Sign-to-911: Emergency Call Service for Sign Language Users with Assistive AR Glasses. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.
[34] Judith A Hall, Terrence G Horgan, and Nora A Murphy. 2019. Nonverbal communication. *Annual review of psychology* 70, 1 (2019), 271–294.
[35] Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*. 588–602.
[36] Yuncheng Hua, Zhuang Li, Linhao Luo, Kadek Ananta Satriadi, Tao Feng, Haolan Zhan, Lizhen Qu, Suraj Sharma, Ingrid Zukerman, Zhaleh Semnani-Azad, et al. 2024. Sadas: A dialogue assistant system towards remediating norm violations in bilingual socio-cultural conversations. *arXiv preprint arXiv:2402.01736* (2024).
[37] Yuncheng Hua, Lizhen Qu, and Gholamreza Haffari. 2024. Assistive Large Language Model Agents for Socially-Aware Negotiation Dialogues. *arXiv preprint arXiv:2402.01737* (2024).
[38] Apple Inc. 2024. Apple Vision Pro. https://www.apple.com/apple-vision-pro/ Mixed-reality headset by Apple Inc., announced in 2023, offering advanced augmented and virtual reality experiences.
[39] INMO Glass. 2024. INMO Air2 - Next-Gen Wireless AR Glasses. https://air2.inmoglass.com/ Accessed: 2024-10-31.
[40] Pegah Jandaghi, Xianghai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2024. Faithful Persona-based Conversational Dataset Generation with Large Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 15245–15270. https://doi.org/10.18653/v1/2024.findings-acl.904
[41] JiWoong Jang, Sanika Moharana, Patrick Carrington, and Andrew Begel. 2024. "It's the only thing I can trust": Envisioning Large Language Model Use by Autistic Workers for Communication Assistance. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.
[42] Hao Jiang, Calvin Murdock, and Vamsi Krishna Ithapu. 2022. Egocentric deep multi-channel audio-visual active speaker localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10544–10552.
[43] Julie Jiang and Emilio Ferrara. 2023. Social-LLM: Modeling User Behavior at Scale using Language Models and Social Network Data. *arXiv preprint arXiv:2401.00893* (2023).
[44] Xiaoqing Jing, Chun Yu, Kun Yue, Liangyou Lu, Nan Gao, Weinan Shi, Mingshan Zhang, Ruolin Wang, and Yuanchun Shi. 2024. AngleSizer: Enhancing Spatial Scale Perception for the Visually Impaired with an Interactive Smartphone Assistant. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–31.
[45] Neha U Keshav, Joseph P Salisbury, Arshya Vahabzadeh, and Ned T Sahin. 2017. Social communication coaching smartglasses: Well tolerated in a diverse sample of children and adults with autism. *JMIR mHealth and uHealth* 5, 9 (2017), e8534.
[46] Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. Prosocialdialog: A prosocial backbone for conversational agents. *arXiv preprint arXiv:2205.12688* (2022).
[47] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
[48] Huining Li, Chenhan Xu, Aditya Singh Rathore, Zhengxiong Li, Hanbin Zhang, Chen Song, Kun Wang, Lu Su, Feng Lin, Kui Ren, et al. 2020. Vocalprint: exploring a resilient and secure voice authentication via mmwave biometric interrogation. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 312–325.
[49] Jiaxing Li, Chi Xu, Feng Wang, Isaac M von Riedemann, Cong Zhang, and Jiangchuan Liu. 2024. SCALM: Towards Semantic Caching for Automated Chat Services with Large Language Models. *arXiv preprint arXiv:2406.00025* (2024).
[50] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957* (2017).
[51] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).
[52] Jiachen Liu, Zhiyu Wu, Jae-Won Chung, Fan Lai, Myungjin Lee, and Mosharaf Chowdhury. 2024. Andes: Defining and Enhancing Quality-of-Experience in LLM-Based Text Streaming Services. *arXiv preprint arXiv:2404.16283* (2024).

[53] Kaiwei Liu, Bufang Yang, Lilin Xu, Yunqi Guo, Neiwen Ling, Zhihe Zhao, Guoliang Xing, Xian Shuai, Xiaozhe Ren, Xin Jiang, et al. 2024. Tasking Heterogeneous Sensor Systems with LLMs. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*. 901–902.

[54] Chaquopy LLC. 2024. Chaquopy: the Python SDK for Android. https://github.com/chaquo/chaquopy. Accessed: 2024-10-30.

[55] Cheng Charles Ma, Kevin Hyekang Joo, Alexandria K Vail, Sunreeta Bhattacharya, Álvaro Fernández García, Kailana Baker-Matsuoka, Sheryl Mathew, Lori L Holt, and Fernando De la Torre. 2024. Multimodal Fusion with LLMs for Engagement Prediction in Natural Conversation. *arXiv preprint arXiv:2409.09135* (2024).

[56] Meta. 2024. Introducing Orion, Our First True Augmented Reality Glasses. https://about.fb.com/news/2024/09/introducing-orion-our-first-true-augmented-reality-glasses/ Accessed: 2024-10-31.

[57] Antje S Meyer. 2023. Timing in conversation. *Journal of Cognition* 6, 1 (2023).

[58] Microsoft. 2024. Limited Access to Speaker Recognition. https://learn.microsoft.com/en-us/legal/cognitive-services/speech-service/speaker-recognition/limited-access-speaker-recognition#registration-process.

[59] Microsoft Learn. 2024. Speaker Recognition Overview. https://learn.microsoft.com/en-us/azure/ai-services/speech-service/speaker-recognition-overview Accessed: 2024-10-31.

[60] Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2023. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. *arXiv preprint arXiv:2311.09180* (2023).

[61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830. https://github.com/scikit-learn/scikit-learn Accessed: 2024-10-30.

[62] Even Realities. 2024. G1: Next-Gen Smart Glasses with Display. https://www.evenrealities.com/g1 Accessed: 2024-10-31.

[63] N Reimers. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084* (2019).

[64] Frank Seide, Morrie Doulaty, Yangyang Shi, Yashesh Gaur, Junteng Jia, and Chunyang Wu. 2024. Speech ReaLLM–Real-time Streaming Speech Recognition with Multimodal LLMs by Teaching the Flow of Time. *arXiv preprint arXiv:2406.09569* (2024).

[65] WIRED Staff. 2024. XRAI Glass Caption AR Glasses: First Look. https://www.wired.com/story/xrai-glass-caption-ar-glasses-first-look/ Accessed: 2024-10-31.

[66] Emma M Templeton, Luke J Chang, Elizabeth A Reynolds, Marie D Cone LeBeaumont, and Thalia Wheatley. 2022. Fast response times signal social connection in conversation. *Proceedings of the National Academy of Sciences* 119, 4 (2022), e2116915119.

[67] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[68] Chongyang Wang, Yuan Feng, Lingxiao Zhong, Siyi Zhu, Chi Zhang, Siqi Zheng, Chen Liang, Yuntao Wang, Chengqi He, Chun Yu, et al. 2024. UbiPhysio: Support Daily Functioning, Fitness, and Rehabilitation with Action Understanding and Feedback in Natural Language. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (2024), 1–27.

[69] 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.

[70] Yu Wu, Zhoujun Li, Wei Wu, and Ming Zhou. 2018. Response selection with topic clues for retrieval-based chatbots. *Neurocomputing* 316 (2018), 251–261.

[71] Zhifei Xie and Changqiao Wu. 2024. Mini-Omni: Language Models Can Hear, Talk While Thinking in Streaming. *arXiv preprint arXiv:2408.16725* (2024).

[72] Lilin Xu, Keyi Wang, Chaojie Gu, Xiuzhen Guo, Shibo He, and Jiming Chen. 2024. GesturePrint: Enabling User Identification for mmWave-based Gesture Recognition Systems. In *2024 IEEE 44th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 1074–1085.

[73] Mengwei Xu, Mengze Zhu, Yunxin Liu, Felix Xiaozhu Lin, and Xuanzhe Liu. 2018. Deepcache: Principled cache for mobile deep vision. In *Proceedings of the 24th annual international conference on mobile computing and networking*. 129–144.

[74] Qianli Xu, Shue Ching Chia, Bappaditya Mandal, Liyuan Li, Joo-Hwee Lim, Michal Akira Mukawa, and Cheston Tan. 2016. SocioGlass: social interaction assistance with face recognition on google glass. *Scientific Phone Apps and Mobile Devices* 2 (2016), 1–4.

[75] Zhenyu Xu, Hailin Xu, Zhouyang Lu, Yingying Zhao, Rui Zhu, Yujiang Wang, Mingzhi Dong, Yuhu Chang, Qin Lv, Robert P Dick, et al. 2024. Can Large Language Models Be Good Companions? An LLM-Based Eyewear System with Conversational Common Ground. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–41.

[76] Bufang Yang, Lixing He, Neiwen Ling, Zhenyu Yan, Guoliang Xing, Xian Shuai, Xiaozhe Ren, and Xin Jiang. 2023. Edgefm: Leveraging foundation model for open-set learning on the edge. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*. 111–124.

[77] Bufang Yang, Lixing He, Kaiwei Liu, and Zhenyu Yan. 2024. VIAssist: Adapting Multi-modal Large Language Models for Users with Visual Impairments. *arXiv preprint arXiv:2404.02508* (2024).

[78] Bufang Yang, Siyang Jiang, Lilin Xu, Kaiwei Liu, Hai Li, Guoliang Xing, Hongkai Chen, Xiaofan Jiang, and Zhenyu Yan. 2024. DrHouse: An LLM-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 4 (2024), 1–29.

[79] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. 2024. Talk2Care: An LLM-based Voice Assistant for Communication between Healthcare Providers and Older Adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–35.

[80] Yao Yao, Zuchao Li, and Hai Zhao. 2024. SirLLM: Streaming infinite retentive LLM. *arXiv preprint arXiv:2405.12528* (2024).

[81] Haolan Zhan, Zhuang Li, Xiaoxi Kang, Tao Feng, Yuncheng Hua, Lizhen Qu, Yi Ying, Mei Rianto Chandra, Kelly Rosalin, Jureynolds Jureynolds, et al. 2024. RENOVI: A Benchmark Towards Remediating Norm Violations in Socio-Cultural Conversations. In *Findings of the Association for Computational Linguistics: NAACL 2024*. 3104–3117.

[82] Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma, et al. 2023. Socialdial: A benchmark for socially-aware dialogue systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2712–2722.

[83] Haolan Zhan, Yufei Wang, Zhuang Li, Tao Feng, Yuncheng Hua, Suraj Sharma, Lizhen Qu, Zhaleh Semnani Azad, Ingrid Zukerman, and Reza Haf. 2024. Let's Negotiate! A Survey of Negotiation Dialogue Systems. In *Findings of the Association for Computational Linguistics: EACL 2024*. 2019–2031.

[84] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2023), 46595–46623.

A Appendix: Prompt Settings

Prompt Template of User Agent

 # OVERALL INSTRUCTIONS
You are playing the role of a speaker, engaging in a conversation with a partner. Refer to the following dialogue and role-play as User 1 to initiate the conversation. Generate your utterance directly without any additional words.
USER PROFILE: $[***Profile***]$
DIALOGUE: $[***Dialogue***]$
When someone approaches to communicate with you, start the conversation naturally and concisely. Keep your responses brief and conversational, similar to everyday interactions, and limit your remarks to 2-3 sentences.

Prompt Template of Conversational Partner Agent

 # OVERALL INSTRUCTIONS
You are playing the role of a speaker, engaging in a conversation with a partner. Refer to the following dialogue and role-play as User 2 to initiate the conversation. Generate your utterance directly without any additional words.
USER PROFILE: $[***Profile***]$
DIALOGUE: $[***Dialogue***]$
When someone approaches to communicate with you, start the conversation naturally and concisely. Keep your responses brief and conversational, similar to everyday interactions, and limit your remarks to 2-3 sentences.

Fig. 22. Prompt templates of the user agent and the conversational partner agent in the dialogue-based role-play.

Prompt Template of User Agent

 # OVERALL INSTRUCTIONS
You are playing the role of a speaker, engaging in a conversation with a partner. Refer to the following dialogue and role-play as User 1 to initiate the conversation. Generate your utterance directly without any additional words.
USER PROFILE: $[***Profile***]$
SOCIAL FACTORS: $[***Social factors***]$
When someone approaches to communicate with you, start the conversation naturally and concisely. Keep your responses brief and conversational, similar to everyday interactions, and limit your remarks to 2-3 sentences.

Prompt Template of Conversational Partner Agent

 # OVERALL INSTRUCTIONS
You are playing the role of a speaker, engaging in a conversation with a partner. Refer to the following dialogue and role-play as User 2 to initiate the conversation. Generate your utterance directly without any additional words.
USER PROFILE: $[***Profile***]$
SOCIAL FACTORS: $[***Social factors***]$
When someone approaches to communicate with you, start the conversation naturally and concisely. Keep your responses brief and conversational, similar to everyday interactions, and limit your remarks to 2-3 sentences.

Fig. 23. Prompt templates of the user agent and the conversational partner agent in the social factor-based role-play.

Prompt Template

 # OVERALL INSTRUCTIONS
You are playing the role of a social assistant, helping a user during live social conversations. Your user is currently engaged in a conversation with a conversational partner. Your role is to provide social suggestions by leveraging the ongoing conversation context, social factors, the implicit personas of both the user and the partner, partner's nonverbal cues, and relevant external information such as weather and news. Your goal is to help the user to initiate and maintain engaging social conversations.

 # TASK INSTRUCTIONS
- Provide insightful social suggestions during conversations.
- Pay attention to the other person's nonverbal cues when generate social suggestions.
- Generate personalized social suggestions considering both the user's and conversational partner's persona cues.
- Keep suggestions brief and easy to understand. First, summarize in 2-3 bullet points without explanation. Save additional tips for the next round. Then, give one simple example sentence.
- You may get an incomplete statement from the conversational partner. You need to understand their intention from these incomplete sentences and provide social suggestions for the user.
- Let's think a bit step by step and Limit your total response to $[***N***]$ words or less.

 # NONVERBAL CUES DEFINIATION
Pay attention to the conversation partner's nonverbal cues during social conversations:
$[***Nonverbal Cues Definition***]$, $[***Nonverbal Cues Guidelines***]$

 # FEW-SHOT DEMONSTRATIONS
Please generate social suggestions referring to the following examples: $[***Few-shot Demonstrations***]$

 # INFORMATION
- Social Factors: Social Norm: $[***Social Norm***]$. Social Relation: $[***Social Relation***]$. Location: $[***Location***]$. Formality: $[***Formality***]$.
- The User's Persona Cues: $[***User Persona Cues***]$.
- The Conversation Partner's Persona Cues $[***Partner Persona Cues***]$.
- Current Nonverbal Cues: $[***Nonverbal Cues***]$.
- External Tools: $[***Weather***]$, $[***News***]$.

Fig. 24. Prompt template of SocialMind.

Prompt of Implicit Persona Extraction

 # OVERALL INSTRUCTIONS

You are provided with several rounds of social conversations between users and their conversational partners. Carefully read these conversations and extract the implicit persona cues.

 # TASK INSTRUCTIONS

- Implicit persona cues refer to personal backgrounds, life experiences, and personal interests.
- Only output the persona cues. If there are multiple cues, separate them with new lines.
- Summarize cues based on all the conversations, rather than individually for each conversation.

 # FEW-SHOT DEMONSTRATIONS

Examples :

$[***Conversation***]$, $[***User Persona Cues***]$, $[***Conversational Partner Persona Cues***]$.

The following are the historical conversations. Now please generate the persona cues of the user and the conversational partner: $[***Social Conversation***]$.

Fig. 25. Prompt of implicit persona extraction in SocialMind.

Prompt Template of Zero-shot

 # OVERALL INSTRUCTIONS

You are playing the role of a social assistant, helping a user during live social conversations. Your user is currently engaged in a conversation with another individual. Your goal is to help the user to initiate and maintain engaging social conversations. Limit your total response to $[***N***]$ words or less.

Fig. 26. Prompt of Zero-shot baseline approach.

Prompt Template of CoT

 # OVERALL INSTRUCTIONS

You are playing the role of a social assistant, helping a user during live social conversations. Your user is currently engaged in a conversation with another individual. Your goal is to help the user to initiate and maintain engaging social conversations. Let's think step by step and limit your total response to $[***N***]$ words or less.

Fig. 27. Prompt of CoT baseline approach.

Prompt Template for LLM Evaluation

 # OVERALL INSTRUCTIONS

You are a strict evaluator responsible for assessing social suggestions during live social conversations. You will review a social conversation involving the user, the conversational partner, and the suggestions provided by the assistant to the user. Focus only on the social suggestion's content and evaluate it based on the following dimensions:

- Personalization. This metric evaluates the quality of social suggestions from a personalized perspective, assessing whether the social suggestions incorporate users' implicit personas, including personal interests and backgrounds. A higher personalization score for social suggestions can enhance user engagement in social interactions, as users are more familiar with the background and topics.
- Engagement. this metric assesses whether the social suggestions consider the conversational partner implicit personas. A higher engagement score can enhance the partner's engagement in social interactions, as they are more familiar with the background and topics.
- Nonverbal Cues Utilization: this metric assesses whether social suggestions take into account the conversational partner's nonverbal cues. For instance, does the social assistant advise the user to observe the conversational partner's facial expressions, such as confusion or frowning, and consider social distance?

Rate the social suggestions on a scale of 1 to 100. Focus solely on the quality of the social suggestions in the dialogue. Additionally, include an explanation for your rating.

 # FEW-SHOT DEMONSTRATIONS

Refer to the following examples: $[***Few-shot Demonstrations***]$

 # GROUND TRUTH

- User personas: $[***User Personas***]$.
- Conversational partner personas: $[***Partner Personas***]$.

Now, let's think step by step and generate the score.

Social conversations: $[***Dialogues***]$

Fig. 28. Prompt template of the LLM evaluation.