Sensible Agent: A Framework for Unobtrusive Interaction with Proactive AR Agents

Geonsun Lee*
University of Maryland
College Park, MD, USA

Min Xia
Google XR
Mountain View, CA, USA

Nels Numan*
University College London
London, UK

Xun Qian
Google XR Labs
Mountain View, CA, USA

David Li
Google XR Labs
Mountain View, CA, USA

Yanhe Chen
Google XR Labs
Mountain View, CA, USA

Achin Kulshrestha
Google XR
Toronto, ON, Canada

Ishan Chatterjee
Google XR
Seattle, WA, USA

Yinda Zhang
Google XR
Mountain View, CA, USA

Dinesh Manocha
University of Maryland
College Park, MD, USA

David Kim
Google XR Labs
Zurich, Switzerland

Ruofei Du†
Google XR Labs
San Francisco, CA, USA

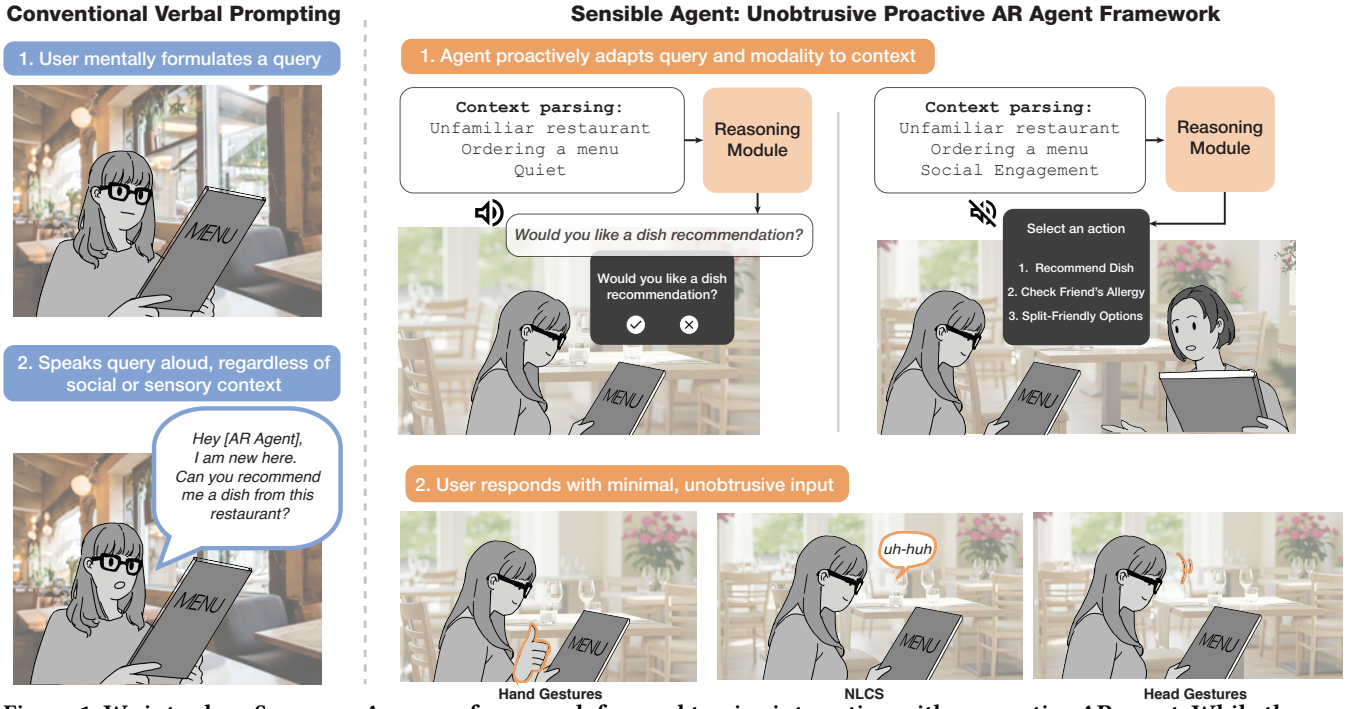

Figure 1: We introduce Sensible Agent, a framework for unobtrusive interaction with a proactive AR agent. While the conventional approach requires users to use voice prompts to instruct agents, Sensible Agent proactively prompts the user based on context, toggles context-adaptive unobtrusive interactions, and suggests different types of queries based on the context.

Abstract

Proactive AR agents promise context-aware assistance, but their interactions often rely on explicit voice prompts or responses, which can be disruptive or socially awkward. We introduce Sensible Agent, a framework designed for unobtrusive interaction with these proactive agents. Sensible Agent dynamically adapts both "what" assistance to offer and, crucially, "how" to deliver it, based on real-time multimodal context sensing. Informed by an expert workshop (n=12) and a data annotation study (n=40), the framework leverages egocentric cameras, multimodal sensing, and Large Multimodal Models (LMMs) to infer context and suggest appropriate actions delivered via minimally intrusive interaction modes. We demonstrate our prototype on an XR headset through a user study (n=10) in both AR and VR scenarios. Results indicate that Sensible Agent significantly reduces perceived interaction effort compared

*This project was undertaken during Geonsun's and Nels's internship at Google.
†Corresponding author: Ruofei Du, me [at] duruofei [dot] com; Also contact: Geonsun Lee, gsunlee [at] umd [dot] edu

UIST '25, Busan, Republic of Korea

ACM ISBN 979-8-4007-2037-6/2025/09
https://doi.org/10.1145/3746059.3747748

to voice-prompted baseline, while maintaining high usability and achieving higher preference.

CCS Concepts

• **Human-centered computing → Mixed / augmented reality**; **Interaction techniques**; *User interface management systems.*

Keywords

Proactive Agents, Augmented Reality, Unobtrusive Interaction, Context-Awareness, Multimodal Interaction, Human-Agent Interaction, Large Multimodal Models, Adaptive Interfaces

ACM Reference Format:

Geonsun Lee, Min Xia, Nels Numan, Xun Qian, David Li, Yanhe Chen, Achin Kulshrestha, Ishan Chatterjee, Yinda Zhang, Dinesh Manocha, David Kim, and Ruofei Du. 2025. Sensible Agent: A Framework for Unobtrusive Interaction with Proactive AR Agents. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25), September 28-October 1, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3746059.3747748

1 Introduction

As augmented reality (AR) and extended reality (XR) technologies become increasingly embedded in everyday life via smart glasses and head-mounted displays, researchers are re-imagining the roles of AI agents. Going beyond simply reacting to users' queries, these agents are envisioned as *proactive assistants*, capable of anticipating and responding to user needs in situ [34, 41]. Users often experience moments of sensory or cognitive disruption, particularly when navigating unfamiliar transit hubs or coordinating actions in crowded public spaces, challenging traditional input-driven interfaces. These moments call for agents that can assess user context and initiate timely support—even when the user's hands, eyes, or voice are unavailable, thereby minimizing the need for explicit interaction.

Recent work has begun to explore proactive agents that surface knowledge or suggestions without direct queries. For instance, AiGet [8] utilizes smart glasses and large multimodal models (LMMs) to deliver incidental knowledge as users explore the world. However, AiGet is primarily designed for curiosity-driven engagement in low-stakes, slow-paced environments. In contrast, many real-world AR use cases involve time-sensitive, socially constrained, or cognitively demanding scenarios where the relevance and delivery of proactive assistance must be carefully tuned. In these settings, the question is not just *what* to suggest, but *whether*, *when*, and *how* it should be delivered. Figure 1 illustrates a motivating example of this contrast: whereas conventional agents rely on explicit voice commands in public, Sensible Agent unobtrusively adapts its prompts and input modalities based on real-time context.

Through a formative study involving 40 participants and 960 context-varying scenarios, we found that user preferences for both content and delivery modality vary widely depending on factors like temporal urgency, environmental sensory load, social presence, and task familiarity. For example, users in loud or crowded environments preferred subtle visual summaries over speech, while those in solo, focused settings accepted more direct audio prompts. These findings highlight the need for systems that treat proactivity as a context-sensitive coordination problem across intent, modality, and timing.

Prior systems address individual parts of this challenge. OmniActions [34] models user activity and context to recommend follow-up actions using LLM-based reasoning, but focuses intent prediction and digital workflow support rather than situated delivery. Human I/O [41] analyzes situational impairments, using multimodal sensing to dynamically adapt interaction modalities, but it does not address proactive intent generation or task-level reasoning. Existing AR interaction techniques often hardcode modality mappings or rely solely on environmental triggers alone [7, 29, 67, 71], limiting their responsiveness to nuanced shifts in user availability, attention, and social context. Instead, we argue that, to be effective, proactive agents must jointly infer both *what* to suggest and *how* to present it, grounded in real-time user context and attentional load. We argue that reasoning jointly over *what* to do and *how* to do it is not simply additive but essential for context-aware AR agents. A well-chosen modality cannot salvage an irrelevant suggestion, and a helpful prompt may go unnoticed if delivered via an ill-suited channel. This integration is particularly critical in socially sensitive or attention-limited settings.

In this paper, we present **Sensible Agent**, a context-aware proactive AR framework that adapts both the content (*what*) and modality (*how*) of its interventions. The framework comprises two modules: (1) an *action suggestion module* using few-shot and chain-of-thought prompting with LMMs to recommend context-relevant agent actions, and (2) a *modality selection module* determining the most suitable delivery channel (audio, visual, gestural, or passive) based on real-time multimodal input like gaze, ambient noise, hand availability. Both modules are guided by a taxonomy of action categories and context variants derived from our study, enabling structured conditioning of LLM outputs and policy decisions.

Sensible Agent is implemented as a WebXR-based prototype and evaluated on a diversity of realistic, daily scenarios presented via an Android XR headset. Our evaluation includes: (1) quantitative annotation of LLM outputs for action and modality appropriateness, (2) real-time latency benchmarking, and (3) scenario-based system demonstrations. We find that our dual-pipeline framework enables proactive support that aligns more closely with user expectations and contextual constraints than single-stream or modality-agnostic alternatives.

In summary, we contribute:

- **Sensible Agent**, a context-aware, proactive AR agent framework that minimizes user interaction effort by jointly determining what to suggest and how to deliver it.
- **A user-derived design implications** of *proactive actions* and *context variants*, collected from a workshop study and a data collection study with 960 user responses across six everyday activities.
- **A functional prototype** using WebXR, multi-modal sensory input (e.g., hand gestures, head gestures, verbal command), and LLM prompting to demonstrate modality- and content-adaptive prompting behavior.
- **A user evaluation** (n=10) quantitatively and qualitatively comparing agent outputs and delivery strategies during a variety of realistic scenarios, presented via XR headset.

Through this work, we take a step toward proactive AR systems that assist unobtrusively, adapt to real-world social and sensory contexts, and reduce the burden of communicating with intelligent agents.

2 Related Work

This section reviews prior work in context-aware AR, proactive agents, and multimodal interaction techniques. These areas collectively inform the design of our framework for unobtrusive, context-sensitive AR assistance.

2.1 Context-Aware AR

Context-aware AR—the ability to perceive and intelligently respond to environmental cues like spatial layout, objects, conditions, and user activity—is fundamental for creating effective and immersive experiences. Recent advances in LLMs have significantly accelerated the capabilities of context-aware AR by enabling richer environmental understanding and more sophisticated spatial reasoning. For instance, Yang et al. [77] conducted a comparative study of Vision-Language Models (VLMs) to evaluate their spatial reasoning capabilities, providing valuable insights for AR applications. Furthermore, XaiR [61], developed by Srinidhi et al., bridges the gap between large multimodal models and XR applications, while Xu et al.'s XAIR framework for multimodal 3D fusion and in-situ learning enables more sophisticated, spatially aware AI interactions [74].

Context-aware AR applications are emerging across various domains. Examples include AI cooking assistants in AR [22, 31], semantic enhancement of object interaction [14, 16, 19], dynamic interface adaptation to context [36], and gaze-based and gesture-based disambiguation [32]. While these AR applications advance interaction with AI, they generally require explicit user input. In contrast, we study unobtrusive, proactive AR, where the system anticipates user needs and offers assistance that requires effortless user interaction.

Commercial XR systems like Apple Vision Pro[1] and Meta Quest Pro[2] offer rich multimodal input such as gaze, gestures, and voice, but their agents remain reactive, relying on user-initiated commands. These systems lack proactive agent behaviors that are contextually modulated based on user state, sensory availability, or social setting. In contrast, Sensible Agent integrates proactive content suggestion with adaptive modality selection, enabling agents to intervene in a timely and socially appropriate manner without requiring explicit user input.

2.2 Agents and Proactivity

Proactive agents aim to enhance user experience by anticipating needs and initiating interactions rather than solely reacting to explicit requests [42, 46, 73, 79, 82]. As outlined in a tutorial paper [39], these proactive behaviors include learning to ask [5, 12, 37, 58, 68, 78, 83, 84], topic shifting [27, 38, 63, 75] and strategy planning with reinforcement learning, counterfactual dialogue act, and label generation [2, 10, 37, 45, 64, 65]. For implementing proactive behaviors, multiple choice question answer [57] allows us to define the problem as next-token prediction aligning well with LLM loss functions and training data.

Beyond these core behaviors, various systems demonstrate proactive capabilities in specific contexts. Parse-Ego4D [1] offers personal action recommendation annotations. Satori [33] proactively guides users by modeling their mental states and environmental context in AR. YETI [3] learns scene understanding for potential intervention. Less or More [69] presents glanceable LLM explanations on smartwatches. COWPILOT supports web navigation [21] by suggesting next steps users can take. Similarly, OmniActions [34] predicts and suggests users action based on multimodal sensory inputs, such as images and audio. The system is triggered by certain actions such as scanning text or the event of taking a picture. While OmniActions focuses on predicting potential user follow-up actions primarily for digital workflow support, our work centers on the challenge of *situated delivery*. Sensible Agent adapts not only "what" assistance to offer, but crucially "how" to present it via *unobtrusive* modalities selected based on real-time multimodal context sensing.

Other research explores proactive engagement for specific user needs. ComPeer is a text-based conversational agent that actively pings users based on previous conversation data and context to provide companionship and mental support [40]. Needs Companion defines a data model for service needs and uses a VA and LLM for needs elicitation and analysis through voice dialogue [48]. Zhang et al. leverages generative agents in a role-playing game to guide users to follow certain actions and in consequence elicit behavioral change such as environment-friendly behaviors [80]. However, these works primarily focus on proactively guiding or intervening with users. In contrast, we center on minimizing interaction friction by adapting "what" assistance to offer and, crucially, "how" to interact based on real-time multimodal context sensing.

2.3 Feedback Channels in Human-Agent Interaction

Effective communication in human-agent interaction relies not only on the agent's output but also on the user's ability to provide appropriate, timely feedback. Prior research has explored a range of modalities through which users can signal feedback, ranging from explicit, intentional utterances to subtle paralinguistic and non-verbal cues.

2.3.1 Explicit feedback. Explicit input methods, such as spoken or typed commands, remain the primary means by which users provide feedback to voice-based agents. Diederich et al. [13] identify communication modality—voice, text, or both—as a central design dimension of conversational agents, enabling users to issue requests, confirmations, or corrections in natural language. Seymour and Van Kleek [59] highlighted the impact of speech as an interaction affordance and described how the shift to conversational interfaces has made interactions with assistants more social in nature.

2.3.2 Whispering. Whispering has emerged as a modality that enables private or low-disruption interactions with voice assistants. Cho [11] examined how whispering affects user perceptions when querying sensitive health information. They found that whispering increased perceptions of social presence and comfort under

[1] Apple Vision Pro: https://www.apple.com/apple-vision-pro/
[2] Meta Quest Pro: https://www.meta.com/quest/quest-pro/

low-sensitivity conditions. Rekimoto [55] introduced DualVoice, a whisper-classification mechanism that distinguishes between whisper and normal speech to support mode switching, e.g., whispering for commands and speaking normally for content input. In a follow-up work, the same author introduced WESPER [56], a system capable of converting whispered speech into audible speech in real-time, enabling silent, unobtrusive interactions in public spaces.

2.3.3 Paralinguistic feedback. Paralinguistic feedback includes non-lexical conversational sounds (NLCS), such as "*mm-hm*", "*uh-huh*", and "*oh*", which convey a range of social cues. Ward [70] identified several categories of such sounds including acknowledgements, affirmation, disagreement, hesitation, and realization. These cues can express user feedback implicitly, signaling engagement, confusion, or alignment and can serve as an unobtrusive input method for users to provide feedback.

2.3.4 Non-verbal feedback. Emerging interaction contexts—such as walking, multitasking, or using devices in public—necessitate alternative input modalities beyond traditional hand or voice input. For instance, Cho et al. [11] examined how users whisper to voice assistants when discussing sensitive health topics, finding that whispering fosters a greater sense of privacy and comfort. Here, we can see that the role of modality is not only in enabling interaction but also in shaping the social acceptability of agent use across contexts.

While walking, traditional interaction techniques such as hand gestures can be unreliable. Zhou et al. [81] found that pinch gestures take significantly longer to execute when users are in motion, and hand input may be unavailable entirely when users are carrying items or wearing gloves. To address this, researchers have investigated hands-free input methods that are less sensitive to body posture or hand availability. These include intraoral input using tongue or lip movement [23], and silent speech recognition through capacitive dental interfaces [25], though these approaches often require specialized hardware.

Other techniques leverage full-body motion during ambulation. Gaze input has been explored for hands-free cursor control, but may divert attention from environmental hazards, introducing safety concerns. In response, several works have explored interaction strategies tailored for mobile AR use. Lages and Bowman [30] proposed interface adaptation techniques for AR transitions during walking. Müller et al. [47] introduced WalkType, which maps lateral walking shifts to interface selection by rendering options as parallel paths on the ground. Kumar et al. [28] extended this technique by combining footpath gestures with gaze for secure AR headset authentication. More recently, GaitGestures [66] demonstrated that intentional changes in stride length and foot strike can be used as a low-effort, hands-free input method during locomotion.

Head-based input has also been explored as a socially acceptable and minimally disruptive modality. Tanenbaum et al. [62] use head rotation to control avatar facial expressions, while commercial systems like AirPods Pro incorporate simple head gestures for call handling. Prior research has also shown that nodding or pointing the nose can enable discrete UI selection[26] and even serve as an authentication mechanism [35], making head gestures a viable modality for subtle, context-aware interactions in AR.

2.4 Interacting with AR in public

Integrating AR devices into public settings presents unique challenges concerning user comfort and social acceptability. Addressing these factors is crucial for facilitating seamless and comfortable public interactions with AR technologies.

Recent studies have explored the social dynamics of AR usage in communal environments. For instance, Kaeder et al. [24] investigated how different virtual display layouts affect users' perceived productivity, feelings of safety, and social acceptability when working with mixed reality in public spaces.

Similarly, Pavanatto et al. [52] examined both user and bystander experiences of XR displays in real-world settings. The study revealed that while users generally accept XR technology in public, factors such as previous XR experience and personality traits can impact perceptions.

Lu and Bowman [43] introduced the concept of Glanceable AR interfaces, designed to provide users with quick, unobtrusive access to information through peripheral displays. Their in-the-wild evaluations demonstrated that such interfaces are less distracting and more socially acceptable for everyday tasks in public settings. Incorporating these insights, our framework emphasizes the development of AR interactions that are not only functional but also socially considerate. By focusing on user-centric design principles, we aim to facilitate AR experiences that users can comfortably and confidently engage with in public settings.

Informed by this prior work on feedback channels, we introduce Sensible Agent. Our framework focuses on unobtrusive proactivity by dynamically adapting both "how" assistance is delivered and "what" feedback modalities are supported, based on real-time multimodal context.

3 Workshop Study

To explore users' motivations and scenarios in which they would require a proactive AR agent, as well as their preferred interaction methods, we conducted a workshop study.

3.1 Study Procedure

We recruited 12 participants internally from Google with diverse backgrounds (engineers, designers, researchers, students). The workshop began by introducing proactive AR agents, contrasting them with the existing user-prompted AR agents (illustrated via a Project Astra [17] video). Using a shared digital whiteboard, participants brainstormed over two structured ideation rounds, each addressing a specific research question:

- **RQ1:** What types of proactive queries would users like the agent to initiate? Specifically, in what *situation* should the agent act, what *action* or *query* should it perform, and *why* is proactive behavior necessary?
- **RQ2:** How should users interact with the agent in public settings? This included considerations of both the *output modality* (how the agent should present its proactive queries) and the *input modality* (how users would respond).

In each round, participants had 10 minutes to reflect and post ideas, followed by group discussion to present and aggregate thoughts. Two moderators ensured equal participation and captured key points. The workshop lasted approximately one hour.

3.2 Findings

We conducted a thematic analysis of responses, drawing on design space frameworks from prior art [44, 76]. We focused on contexts where users desired proactivity, the actions they expected, and their reasoning. These findings shaped our framework's *"what"* and *"how"* modules. These findings shaped the design of our framework, particularly how the agent determines *"what"* to do and *"how"* to present it. Participants are denoted as W1–W12.

3.2.1 ***When*** *do we need proactive AR agents?*

Participants described scenarios in which they expected agents to anticipate needs and act without explicit input. Across 12 participants and 45 scenarios, six recurring contextual factors emerged:

Repetitive, Predictable Activities (n=9). Participants described routine scenarios in which their next actions were both predictable and repetitive. They expressed frustration with having to repeatedly issue explicit commands for tasks they perform frequently. For instance, W1 noted, *"I have daily routines such as taking a bus and playing Spotify. In this case, the agent should learn this and suggest it first when I get on the bus".*

Public or Socially Awkward Situations (n=6). Participants noted that verbal interaction was socially uncomfortable or inappropriate in public or quiet environments (e.g., libraries, cafes).

Uncertainty or Lack of Awareness (n=4). Participants described situations in which they were unsure what assistance to request or unaware of the agent's available capabilities. In these cases, proactive suggestions were seen as beneficial. W9 mentioned: *"Sometimes, I don't even know what's possible to ask; proactive suggestions would help me discover useful actions."*

Unfamiliar Environments or Activities (n=5). Participants desired proactive guidance in new settings or during unfamiliar activities, such as visiting a new city or starting a hobby. Some emphasized the need for varied suggestions in these situations (W6, W7, W10).

Time-sensitive Scenarios (n=4). Participants identified high-pressure contexts (e.g., rushing to catch transport) as prime opportunities for proactive assistance. W5 said: *"When I'm rushing, I don't have the mental capacity to have a full blown conversation with an agent."*

These findings illustrate the necessity of considering contextual variables like familiarity, social environment, urgency, and uncertainty to effectively shape ***what*** proactive actions the agent should suggest.

3.2.2 ***What*** *do we want proactive AR agents to do?*

We found significant overlap between the contexts of ***when*** and participants' suggestions for ***what*** the agent should do. However, participants also explicitly articulated desired proactive actions:

Information Delivery (n=11). Delivering relevant context-aware information without explicit user query (*e.g.*, translating menus, recognizing landmarks, or offering pronunciation feedback in language learning).

Reminders and Notifications (n=9). Nudging users about forgotten intentions or events based on routine or temporal triggers (*e.g.*, picking up medication, sending messages when late, stocking household items).

Suggestion and Option Surfacing (n=6). Offering creative or exploratory ideas when users have no concrete goals (*e.g.*, suggesting interior decor, restaurant options, or AR visualizations for artwork).

Error Detection and Guidance in Tasks (n=5). Providing real-time guidance during procedural or skill-based tasks when users pause, struggle, or deviate (*e.g.*, correcting instrument fingering, helping with furniture assembly, pointing out cooking errors).

Environment or Object Control (n=4). Automatically interacting with physical or digital systems based on routine or context (*e.g.*, turning off lights, logging food, muting phone calls while driving).

While specific actions varied, participants consistently preferred contextually timed, non-intrusive suggestions that reduced the burden of remembering, navigating UIs, or formulating questions.

3.2.3 ***How*** *do we want to interact with proactive AR agents?*

Participants emphasized the need for unobtrusive, contextually appropriate interaction methods:

Hand Gestures (n=10). Participants frequently suggested hand gestures as subtle means of interaction but explicitly noted limitations when engaged in hand-intensive tasks.

Head Gestures (n=5). Simple head movements such as nodding, shaking, or slight tilting were popular due to their subtlety and intuitive nature. *"If the agent asks a simple yes-no question, I could just slightly nod or shake my head without anyone noticing,"* explained a participant (W7).

Gaze-based Interactions (n=5). Gaze inputs like blinking or gaze-dwelling were identified as useful, especially when hands were unavailable or the user wished to interact privately.

Subtle Auditory Inputs (n=4). Participants suggested NLCS and whipsering as an alternative for a less intrusive way than overt speech.

Integrated Activities (n=3). Some participants proposed embedding response to proactive suggestions into ongoing activities (*e.g.*, continuing a task as implicit confirmation).

Nine participants explicitly or implicitly referred to Situationally Induced Impairments and Disabilities (SIIDs) [41, 72], which influenced their preferences for both how the agent should present itself and how they would interact with it.

3.3 Design Implications

Our findings suggest that users' expectations for proactive AR agents are shaped not only by the activity at hand but also by fine-grained situational factors that influence the relevance of proactive actions (*what*) and the appropriateness of interaction modalities (*how*). We outline five key design implications that directly informed our framework:

(1) The same activity may require different proactive behaviors based on contextual variants. Participants often described repeated tasks (*e.g.*, navigating an airport, visiting a museum) where the proactive action varied depending on environmental familiarity, time pressure, or social setting. This suggests that static task-based modeling is insufficient. Proactive systems must account for how variations in context shift the user's expectations.

Our framework addresses this through a context similarity module in the *what* pipeline, allowing the system to adapt actions to nuanced differences across similar scenarios.

(2) Social engagement and public settings significantly constrain interaction modalities. Participants expressed reluctance to interact with agents via speech or overt gestures in socially sensitive environments (*e.g.*, meetings, public transit). This highlights the need to reason about social acceptability—not just sensor availability—when choosing output and input modalities. In our framework, this is addressed by incorporating social engagement as an explicit factor in determining interaction modality.

(3) Temporarily impaired input/output channels are common in everyday settings. Rather than permanent disabilities, participants frequently described moments where they were visually, audibly, or physically unavailable due to the activity (*e.g.*, eating, driving, holding an object). These temporary impairments—also described in prior work such as SSID [41]—should be treated as first-class input to the system. Our framework integrates this insight into the *how* module, enabling dynamic modality selection based on real-time multimodal sensing.

(4) Users welcome suggestion diversity when uncertain, but prefer precision when familiar. When participants were unsure of their goals or facing novel tasks (*e.g.*, decorating a room, visiting a museum abroad), they welcomed diverse suggestions. In contrast, familiar routines called for focused, streamlined actions. In response to this, our system's suggestion generation module varies the breadth and structure of proactive queries (*e.g.*, binary vs. multi-choice) depending on user familiarity with the activity.

(5) Embedded, multimodal confirmations lower friction in high-effort scenarios. Participants often preferred confirming or rejecting proactive suggestions through natural, low-effort behaviors (*e.g.*, nodding, gaze dwelling, continuing the current task). This indicates that confirmation should be implicitly embedded in the interaction rather than handled through explicit follow-ups. Our interaction module prioritizes combining multimodal signals (head, hand, voice, gaze) to enable confirmation mechanisms with minimal cognitive or physical effort.

4 Sensible Agent: A Framework for Context-aware Unobtrusive Proactive Agents

We present **Sensible Agent**, a framework for building proactive AR agents that prioritize unobtrusive interaction and minimal user effort. Unlike traditional systems that rely on user-initiated queries, our framework is designed to anticipate user needs and respond proactively, while adapting both the content and delivery of suggestions to situational context.

The framework consists of two interdependent reasoning modules, as illustrated in Figure 2: the `ACTION RECOMMENDATION MODULE` (*What*) and the `INTERACTION ADAPTION MODULE` (*How*). While our prototype implements the core components of both modules, certain capabilities such as context similarity based on long-term user history are part of the envisioned design and discussed as future directions.

4.1 Action Recommendation Module (What)

This module determines what actions the agent should proactively suggest in a given context. It is designed to anticipate user intent and reduce decision-making burden by tailoring suggestions to situational cues and prior behavior.

User Context Adaption. The `CONTEXT SIMILARITY MODULE`uses the *current user context*, extracted in real time and encompassing dimensions such as activity and location familiarity, perceived cognitive load, social engagement, and temporal urgency. In the framework, it also draws on a *user context history* and *prior user actions*, enabling the system to learn from repeated patterns or behavioral regularities.

LMM Reasoning. Based on the current context—and user history of similar context—the `PROACTIVE ACTION MODULE` not only determines *what* action(s) to suggest but also selects the most appropriate *presentation format*. We define three three primary formats, ordered from most to least effort for the user:

- **Multi-choice Selection:** Presented when several possible actions may be contextually appropriate, allowing the user to choose from.
- **Binary Confirmation:** Employed when a single action is predicted with high confidence, but explicit user confirmation via a 'yes'/'no' respond is required.
- **Icon-based Cue:** Deployed for highly probable, low-stakes actions where user intent is inferred with very high confidence. The agent proactively visualizes a relevant graphical icon (*e.g.*, a translation/menu icon) in peripheral region, affording user interaction with minimal interruption.

4.2 Interaction Adaption Module (How)

This module determines how the proactive suggestion should be delivered and how the user should interact with it, based on real-time input/output availability and context-driven appropriateness.

User Context Adaption with I/O Channels Availability. The `HOW` module shares the same core *user context* extracted for the `WHAT` module but further considers *input-related context*, such as whether the user's hands are occupied, their environment is noisy, or they are engaged in conversation. These additional cues reflect situationally induced constraints that affect interaction feasibility.

User Input Similarity. While not yet implemented, `INTERACTION CONTEXT SIMILARITY MODULE` component would compare current input-related context to prior interaction patterns, helping refine modality decisions based on similarity to previously successful input conditions. This design is informed by the concept of SIIDs [41], which describe temporary constraints on user input/output channels.

Presentation Strategy. The `QUERY PRESENTATION MODULE` selects how the agent's proactive suggestion is presented to the user, choosing from:

- **Visual-only:** On-screen UI elements, icons, or overlays.
- **Auditory-only:** Spoken messages or system voice prompts.
- **Audio-visual:** Redundant or complementary presentation across both channels.

Presentation strategy is determined based on environmental and social context. For example, in a quiet or very noisy public setting,

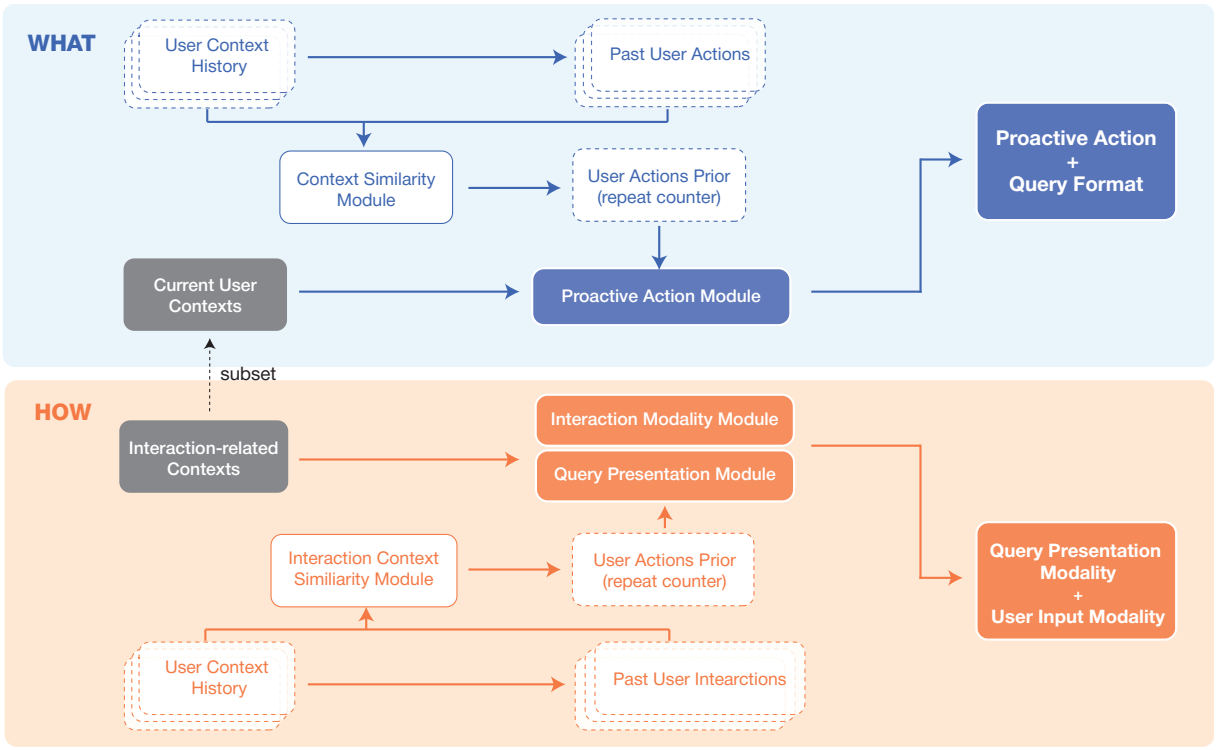

Figure 2: Detailed dataflow of the Sensible Agent framework. An `ACTION RECOMMENDATION MODULE` (`WHAT`) takes user context and determines the suggested action in one of three primary formats, and an `INTERACTION ADAPTION MODULE` (`HOW`) selects presentation modality and input modalities.

the agent may suppress audio and rely on visuals, while in visually demanding tasks (*e.g.*, biking), auditory prompts are prioritized.

Interaction Modality Adaption. The `INTERACTION MODALITY MODULE` determines which input modalities are enabled for confirming or responding to proactive suggestions. It considers both the user's input-related context and the selected presentation strategy. For instance, if the output is visual-only and the user is not looking at the screen, gaze-based input is not viable. Modalities supported include:

- **Gaze (dwell):** Used for visual interfaces, enabling binary or multi-choice input by tracking where the user looks. Buttons are triggered by holding still for one second.
- **Hand gestures:** Uses explicit gesture recognition (*e.g.*, open palm, fist) rather than raycast-based pointing, enabling confirmation or selection even when direct targeting is not feasible.
- **Head gestures:** Supports nodding and shaking for binary prompts, and directional tilting (*e.g.*, left, right, backward) for multi-choice selection.
- **Voice input:** Enables spoken responses using lightweight verbal commands. For binary interactions, users may respond with naturalistic non-conversational lexical sounds (NCLS) such as "*uh-huh*" or "*mmm-mm*." For multi-choice prompts, the agent supports one-word commands such as "*one*," "*two*," or "*three*."

These input modalities can be used independently or in combination, depending on user availability and task constraints. The design prioritizes interaction methods that are socially acceptable and impose minimal cognitive or physical load.

4.3 Module Integration

The two modules operate in parallel and share context inputs. The `WHAT` module determines the structure and content of the suggestion, which then informs the `HOW` module's choice of presentation and interaction modality. For instance, a multi-choice suggestion during a high-cognitive-load task may trigger a visual interface with hand gesture input, whereas an icon-based prompt during a routine task may use gaze-only interaction.

Together, these modules support a context-aware proactive agent that adapts not only to what the user needs, but how they can most easily engage. In the following section, we describe how we implemented the core components of this framework in a functional prototype and outline the system capabilities currently supported.

5 Prototype Implementation

We implemented a WebXR-based working prototype of our context-aware proactive AR agent system, focusing on the core modules identified in our framework: the proactive action module and the adaptive interaction modality module. Our system leverages LMMs

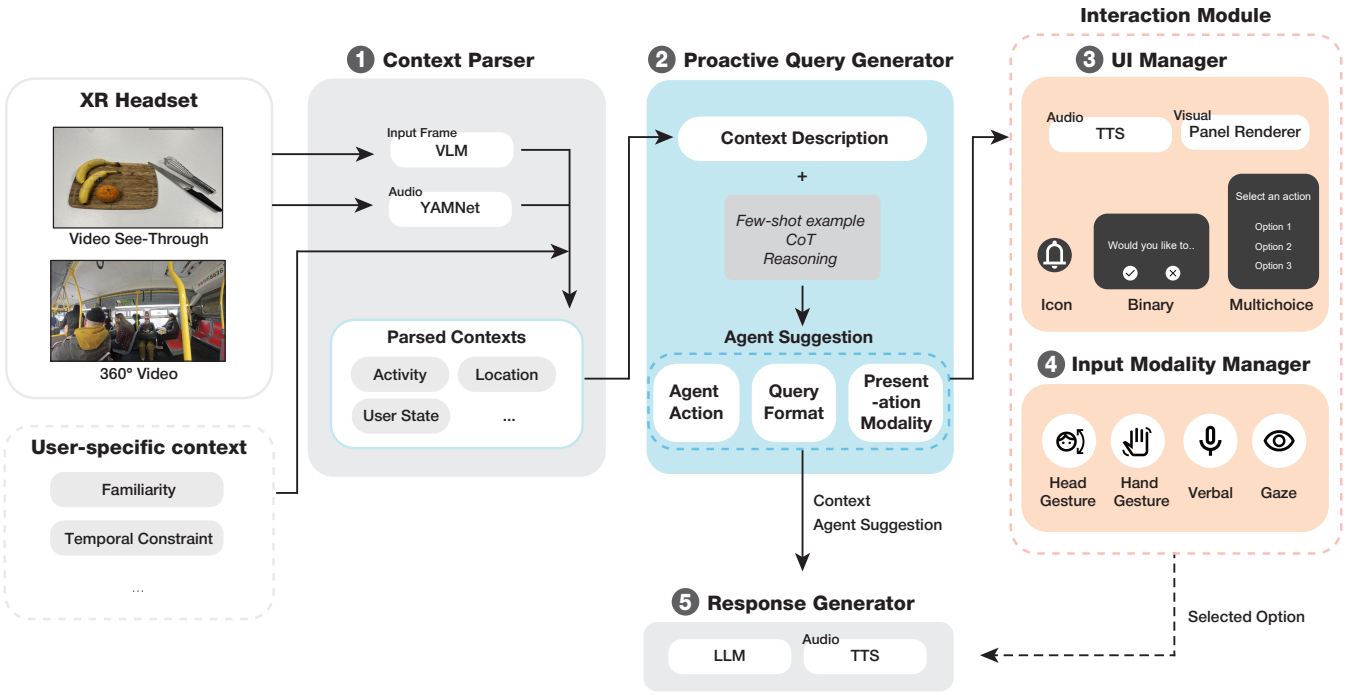

Figure 3: System architecture of our proactive AR agent prototype. The full system is implemented in WebXR with support for real-time interaction in 360° videos or video see-through AR environments. The system processes visual and audio input (1) and parses contextual attributes such as familiarity, urgency, and environmental noise using a VLM and YAMNet. (2) Based on the parsed context, the proactive query generator formulates a suitable suggestion, including its agent action, presentation modality, and query type. These are passed to the interaction module, (3) where the UI manager renders the query and the (4) input modality manager enables one or more input modalities (*e.g.*, gaze, hand, head, voice) based on feasibility and appropriateness. The interaction module then forwards the selected option by the user to the (5) response generator.

to infer real-time context and generate proactive suggestions and interaction strategies accordingly.

We describe the overall architecture and interaction flow of the system, followed by details of our data collection study, which informed the agent's suggestion generation logic.

5.1 System Architecture

Figure 3 illustrates the overall architecture of our prototype, including context parsing, query assembly, inference, and agent response integration. The system operates on egocentric video input, which can come from a real-time video see-through (VST) stream through Android's Camera2 API[3] or a 360° pre-recorded video. The latter is primarily used to simulate environmental contexts in controlled study conditions where live capture is not feasible.

When a trigger event is detected–such as a pause in activity or gaze fixation–a single image frame is extracted and sent to the system's reasoning pipeline. This pipeline consists of three layers: a context parsing layer, a proactive query generation layer, and a response generation layer. Each layer operates using LMMs (GPT-4o), conditioned through in-context learning with structured input examples derived from our data collection study.

[3]Camera2 API: https://developer.android.com/media/camera/camera2

Context Parsing. The first layer uses the visual input along with manually injected user-specific variables, such as task familiarity or temporal urgency, to identify key contextual attributes. These include physical environment (*e.g.*, type of space), user state (*e.g.*, hands occupied, conversational setting), and social or environmental constraints (*e.g.*, noise level, presence of others). This layer simulates the role of the context similarity module in our framework, although it does not perform explicit similarity computation across historical data. We use Chain-of-Thought (CoT) prompting here to extract structured outputs through intermediate reasoning steps, enabling more accurate context decomposition.

Proactive Query Generation. The output from the first layer is passed to the second layer, which generates the proactive query the agent should present. This includes three elements: the action content (*e.g.*, suggesting information or assistance), the query format (multi-choice, binary, or icon-based), and the suggested presentation modality (visual, auditory, or both). Although our framework initially distinguishes between the `WHAT` and `HOW` modules, we found that the decision around presentation modality is closely entangled with the action content and context. As such, this component is computed together with the query in the second layer, while the input modality constraints are handled downstream.

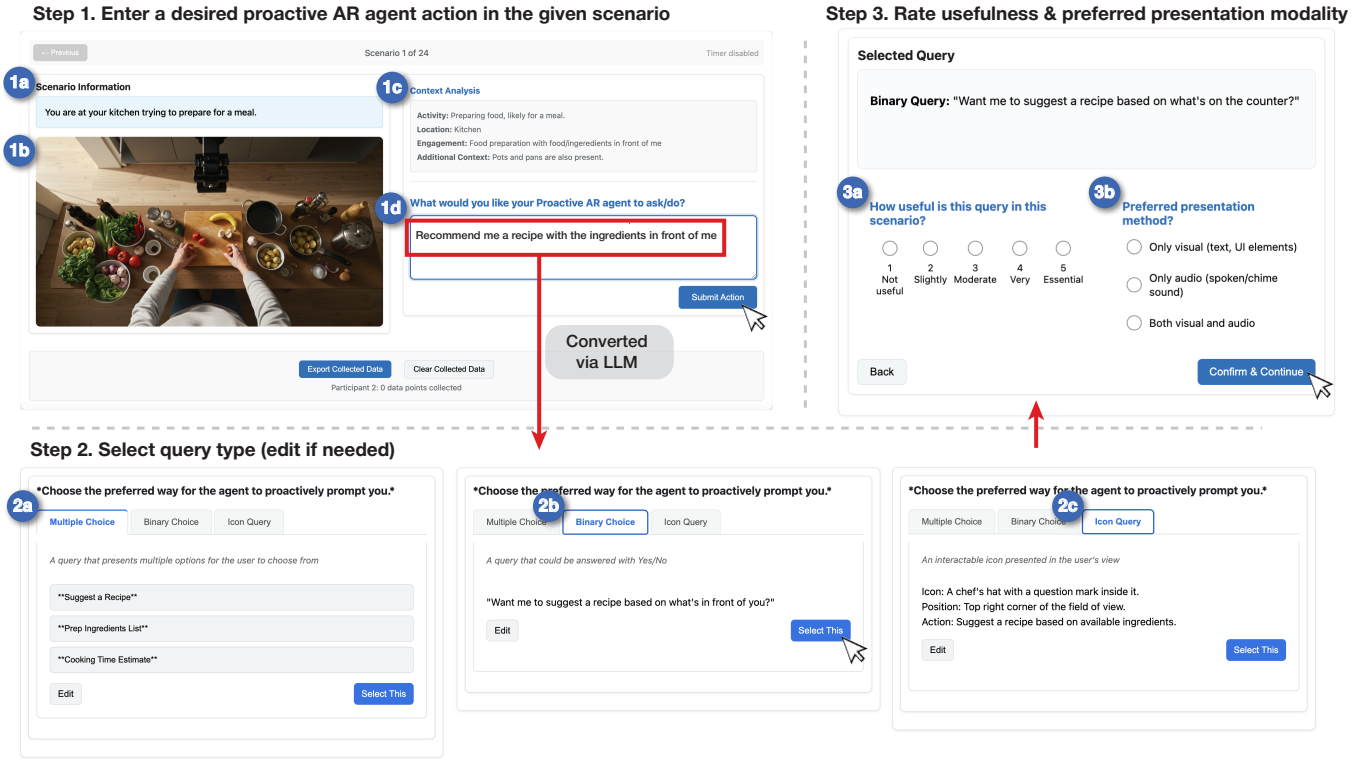

Figure 4: Web interface for the data annotation study. Each participant annotated 24 scenarios through a 3-step workflow. In Step 1, participants viewed: (1a) a short text describing the scenario, (1b) a synthetic egocentric image for visual consistency, (1c) contextual details (*e.g.*, location, engagement), and (1d) a text input field to describe the desired proactive AR agent action. In Step 2, the input was converted into (2a) multi-choice, (2b) binary, and (2c) icon-style queries using LLMs. Participants could edit or choose their preferred query type. In Step 3, they (3a) rated the usefulness of the action and (3b) selected the preferred presentation modality (audio, visual, or both). Final responses were exported as a CSV after completing all 24 scenarios.

Interaction Module. The UI manager then constructs a panel interface or an audio playback using OpenAI's text-to-speech (TTS) API [51] based on the generated suggestion and modality. At the same time, the parsed context is sent to the input modality manager, which enables a subset of input modalities. These include head gestures, hand gestures, gaze (via dwell), and verbal inputs. Each modality is gated based on two criteria: the inferred contextual appropriateness (e.g., SIIDs [41]) and the feasibility given the chosen output modality . For example, if the context suggests that the user is in a noisy public setting, voice input is disabled and visual interactions such as gaze or head gestures are prioritized. Similarly, if a query is presented solely in audio form, gaze interaction is considered infeasible and suppressed.

Response Generation Layer. Once the user confirms or selects an option from the proactive prompt, the system passes both the structured context and the selected action to an LLM to generate a natural language response. This response serves as the agent's follow-up behavior, grounded in the user's selection (e.g., providing details about a painting the user is viewing in a museum). The generated utterance is then synthesized via TTS and played back through the headset's audio channel.

5.2 Proactive Actions Data Collection Study

To support context-aware query generation in our proactive agent system, we conducted a data collection study designed to elicit how users' expectations of agent behavior vary based on situational context. This study informed the core reasoning mechanism of our system by providing grounded examples of context-action-modality mappings, which were later used to condition an LLM through in-context learning.

Web-Based Annotation Interface. We developed a custom web-based annotation interface (see Figure 4) that presents participants with egocentric-view scenarios. Each scenario consisted of a synthetic image paired with a one- to two-sentence description simulating an immersive AR context. To ensure interpretability and consistency, we also provided optional structured descriptors–such as location, detected high-level activity, user engagement, and environmental or social context–displayed in a collapsible panel. These served as objective reference points for participants, complementing the visual and narrative descriptions.

Participants were asked to imagine themselves in each situation and enter, in free-text form, what they would want a proactive AR agent to do or ask on their behalf. Upon submission, an LLM (GPT-4o) was used to reformat the user-described action into three

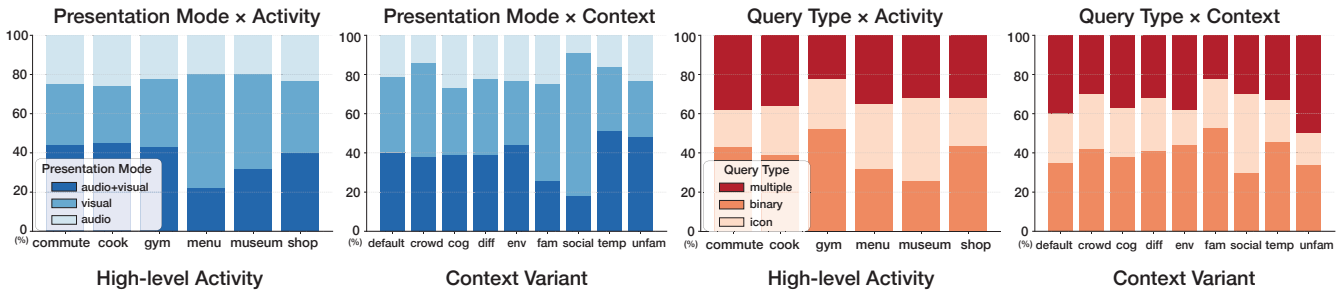

Figure 5: Distribution of data entries in selected presentation modes (left) and query types (right) across different high-level activities and context variants. Data showed varying preferences for modality (audio, visual, audiovisual) and query format (binary, multiple-choice, icon-based) depending on situational demands and activity type.

proactive query forms: multi-choice, binary, and icon-based. Participants could revise the generated phrasing and select the query form they felt was most appropriate for the given context. They were also asked to rate how useful they believed the proactive suggestion would be in the scenario on a five-point Likert scale, and to specify their preferred modality of presentation (audio, visual, or both). Participants could navigate between scenarios at any point and revise their annotations.

Presented Scenarios. The study included 48 scenarios, each representing a variant of one of six high-level activities commonly encountered in daily life: reading a menu at a restaurant, working out at a gym, grocery shopping, browsing in a museum, commuting by public transportation, and cooking in a kitchen. Each activity contained 5 to 8 contextually distinct variants, which were designed to reflect factors such as location and activity familiarity, situational impairments (*e.g.*, hands occupied), social constraints, and temporal urgency. Each participant annotated 24 scenario of three selected high-level activities, with the task taking approximately 30 to 40 minutes to complete.

Participants. We recruited 40 participants through internal mailing lists and social platforms. Participants varied in age (21 - 44) and background and included individuals with prior experience using AR headsets or glasses (34 out of 40). On average, participants reported a $\mu = 3.97$ mean experience with AR and $\mu = 3.57$ with voice assistants on a 5-point Likert scale.

5.2.1 Results and Analysis. We analyzed 960 query entries from 40 participants (24 per person) across six high-level activities and their contextual variants. Our analysis focuses on (1) consistency in query preferences, (2) variation in query type and presentation modality by context, and (3) a taxonomy of desired proactive actions. (Figure 5)

Dataset Overview We analyzed a total of 937 proactive action entries from 40 participants across 48 scenarios. Each participant encountered scenarios drawn from six high-level activity types—menu reading, cooking, visiting a museum, commuting, working out, and grocery shopping—embedded in varied contextual conditions such as familiarity, temporal urgency, or social engagement. For each scenario, participants (1) described their desired agent action, (2) selected a preferred query type (binary, multi-choice, or icon), (3) chose a presentation modality (audio, visual, or audio+visual), and (4) rated the usefulness of the action out of a 5-point Likert scale. To understand preferences for *how* the agent should assist, this analysis focuses on the 937 of 960 entries (97.6%) where users desired proactive help (usefulness rating > 1).

Query Format is Shaped by Contextual Demands Query type selection was highly context-sensitive. Overall, multi-choice (42%) and binary queries (36%) were more frequently preferred than icon-based options (22%) across high-level activities. However, preferences shifted meaningfully across context variants.

In unfamiliar scenarios, users often favored multi-choice formats to explore alternative paths or receive richer input from the system (50%)). For example, in an unfamiliar restaurant context, a participant selected a multi-choice query, *"1.Translate dish names, 2.show images, 3.suggest the most popular dish,"* and rated it highly useful (5/5). On the other hand, binary queries were more common under temporal pressure or when rapid assistance was needed (48%)), such as, *"Would you like me to recite your grocery list?"*

Icon-based formats, though less common overall, emerged in socially sensitive (40%)) or familiar environments (25%) where minimal interaction was desired. In one such instance, a participant requested, *"Show vegan options for my friend,"* during a socially-engaged dining setting, paired with an icon query and visual-only presentation.

These findings suggest that the agent's querying mechanism should be sensitive to both task complexity and situational constraints, offering lower-friction formats in fast-paced or public-facing contexts while enabling richer interactions when users have time and attention to spare.

Presentation Modality Vary by Contextual Constraints Across all scenarios, the majority of participants preferred audio-visual presentation (38%), likely due to the redundancy and clarity it offers. However, this preference was not universal.

In socially dense or quiet public settings, such as museums or restaurants, users gravitated toward visual-only queries. For example, in a crowded museum, one user asked the agent to *"Show artwork info I am looking at,"* choosing visual-only presentation for discretion. On the other hand, audio-visual modalities were favored in unfamiliar or time-constrained scenarios, where rapid and clear communication was necessary—*e.g., "Suggest fast options I can eat from the menu."*

These results reinforce the need for modality adaptation in agent design, modulated by situational context and the user's social environment.

Context Description	CoT Reasoning	Agent Action	Query Type	Modality
User is in a museum, crowded with people and slightly noisy, while engaged with an art piece.	User may not hear audio clearly and is visually focused on the artwork. A visual, low-effort query is ideal.	Offer more information about the artwork (e.g., title, artist, background).	Icon	Visual
User is in a familiar grocery store but is in a rush, quickly moving through aisles.	User's gaze is shifting frequently; visual queries may be missed. Audio is preferred.	Offer to recite the user's grocery list	Binary	Audio
User is alone in a new restaurant, unfamiliar with the menu. The space is quiet and not crowded.	User may need help deciding what to order and may benefit from both visual and audio support.	Provide dish recommendations (*e.g.*, "Top dishes," "Vegetarian options," "What I had last time").	Multi-choice	Audio + Visual

Table 1: Representative Few-Shot Examples for Context-Conditioned Query Reasoning

Query Format Stability within Task Categories We analyzed query type consistency across contextual variants of each activity. A participant's query type was considered consistent for a task if it appeared in at least 80% of the contextual variants for that activity. Out of 240 possible activity-participant pairs, 23 met this criterion.

This relatively low consistency rate indicates that users do not adopt a one-size-fits-all querying strategy. Instead, they fluidly adjust their preferences depending on the situation—supporting the notion that context-aware query adaptation is critical for proactive systems.

Taxonomy of Desired Proactive Agent Actions To systematically analyze participants' free-text responses, we developed a two-layer taxonomy.

At the first level, we identified core **action categories**, including *Suggest*, *Remind*, *Guide*, *Summarize*, *Automate*, *Visual Augmentation*, *Information Retrieval*, and *Take App Action*. For instance, a participant in a cooking scenario requested, *"Detect my step and display the next one"*, which was categorized as a *Guide* action.

At the second level, we derived **contextual categories** that modulate these actions, such as *Familiarity-Based*, *Urgency-Based*, *Social Coordination*, *Cognitive Load*, and *Sensory Disruption*. For example, one participant in a restaurant setting asked the agent to *"brighten the menu and read items aloud,"* which combines *Visual Augmentation* with *Sensory Disruption*.

A list of context categories and action types as well as the distribution among annotated data is shown in Appendix A. This analysis provided a structured understanding of user preferences, which in turn guided our manual authoring of the representative few-shot prompts used for in-context learning.

5.2.2 Context-Conditioned Query Reasoning with In-Context Learning. To generate context-appropriate proactive agent behavior, we implemented an in-context learning pipeline using LLM. Rather than relying on fixed rules or templates, this module adapts to novel situations by conditioning on structured representations of user context. It outputs a proactive agent action, query type, and presentation modality. The querying strategy draws on our dataset of 937 annotated examples and our taxonomy of user preferences (§5.2.1).

Prompting Strategy. Each input to the language model consists of three components: (1) a preamble specifying the agent role and the valid output space; (2) a small set of few-shot examples drawn from our annotated pool; and (3) the user's current task context, expressed as natural language with key situational factors (activity, location, sensory load, social engagement, familiarity, urgency). Few-shot examples are composed of three fields:

- **Context:** A natural language description combining relevant situational dimensions such as high-level activity (*e.g.*, cooking, commuting), physical location (*e.g.*, kitchen, museum), environmental factors (*e.g.*, noise, crowd density), user engagement (*e.g.*, hands occupied, visually focused), social context (*e.g.*, alone or with others), and task familiarity or urgency (*e.g.*, rushing, first-time visit).
- **Reasoning (Chain-of-Thought):** A brief rationale that connects salient context features to interaction considerations—*e.g.*, decision complexity, input/output availability, or social norms. These reasoning lines were manually authored based on design patterns extracted from our user study, and serve as Chain-of-Thought (CoT) scaffolds for the model.
- **Agent Suggestion:** A structured prediction consisting of (a) a description of the agent's suggested action, (b) a query format (binary, multi-choice, or icon), and (c) a presentation modality (audio, visual, or audio+visual).

For instance, a complete triplet might include:

- *Context:* "User is in a grocery store, browsing aisles alone, holding a shopping cart, and navigating quickly in a noisy, crowded setting."
- *Reasoning:* "Because the user is rushing and both visually and physically engaged, a binary audio prompt reduces interaction load."
- *Agent Suggestion:* Offer to recite the user's grocery list | Binary | Audio

Table 1 shows representative exemplars and how context cues map to different query policies.

Few-Shot Selection. We provide six exemplars, selected for contextual similarity—matching on high-level activity and at least one additional contextual category (e.g., SIID, social engagement, or familiarity). This balanced diversity with input-length constraints. The prompt then concludes with the user's current context, followed by the final task instruction:

```
Based on the context provided above, generate:
(1) a reasoning for your decision,
(2) the recommended agent action,
(3) a query format (binary/multi-choice/icon), and
(4) a presentation modality (audio/visual/audio+visual).
Structure the output as shown in the examples; if
the query format is 'multi-choice', provide three
distinct options for the agent action.
```

Runtime Inference and Output Integration. At runtime, a lightweight parser composes a context string from scenario tags or simulated sensor values (e.g., engagement, noise level). The LLM returns an *agent action*, *query type*, and *presentation modality*. We apply simple string parsing to extract these fields and pass them to the interface for immediate rendering (Figure 3). This produces differentiated behavior across context variants while keeping latency within our system constraints (Appendix B).

5.3 Interaction Modality Implementation

To support unobtrusive interaction in a variety of contexts, we implemented four user input modalities for responding to the agent: voice, head gestures, hand gestures, and gaze. All interaction methods were developed using WebXR[4] and three.js[5], and run in Chrome v137 on an Android XR headset.

5.3.1 Verbal Interaction. We implemented verbal input using the Web Speech API, specifically the 'SpeechRecognition' interface. The recognition system operates in a limited vocabulary mode, accepting discrete responses such as `YES`, `NO`, `ONE`, `TWO`, and `THREE`, corresponding to binary and multi-choice prompts. To improve recognition accuracy and avoid partial matches within longer phrases, we applied regular expression boundaries (*e.g.*, `\bONE\b`). Recognition is filtered by a confidence threshold of 0.7 to reduce false positives.

To support NLCS interaction, we trained a lightweight, real-time classifier using Google's Teachable Machine [9]. The classifier distinguishes ambient background noise from affirmative and negative NLCS signals. We trained the model using 120 samples of background noise and 30 samples each of affirmative and negative humming. Internally, Teachable Machine uses a transfer-learned version of Google's YamNet audio classification model, which employs the MobileNet architecture [20] and operates on mel spectrogram representations. The trained model was exported to TensorFlow.js [60] and integrated into our WebXR-based framework for real-time NLCS detection.

5.3.2 Head Gestures. Head-based interaction leverages the user's rotational head pose, estimated from WebXR's head tracking data. For binary queries, we detect nodding (pitch axis) and shaking (yaw axis) based on oscillatory movement patterns. A gesture is confirmed when 3–4 directional reversals are detected within a fixed temporal window, with a per-frame angular velocity threshold of 0.05 radians to filter out noise and micro-movements.

To enable multi-choice selection, we implemented a tilt-based gesture mapping. This tracks the user's absolute head orientation relative to a neutral reference pose. Tilts beyond 0.3 radians to the left or right are interpreted as selecting the first or second option, respectively. A backward tilt exceeding 0.4 radians selects the third option. This mapping provides a hands-free alternative to UI selection, while maintaining a low motor demand footprint (e.g., when user is washing hands).

5.3.3 Hand Gestures. Hand input is recognized through WebXR's joint tracking API and analyzed using a geometry-based classifier. Our implementation defines five discrete hand gestures: `ONE`, `TWO`, `THREE`, `THUMBS_UP`, and `THUMBS_DOWN`. Gestures are recognized based on the extension state of individual fingers, computed via vector alignment and extension ratios derived from the relative positions of joint triplets on each finger.

A finger is marked as extended if the dot product between its segment vectors is near 1 (indicating alignment) and the overall extension exceeds 80% of its normalized length. For example, a `TWO` gesture is recognized when the index and middle fingers are both extended, while others remain curled. Thumbs up and down are detected via the dot product between the thumb vector and the global up/down direction, based on the palm normal.

To prevent false positives from brief postural noise, gestures must be held for a minimum duration of 1000ms, sampled at 30ms intervals. Gesture state is tracked over time to ensure stable classification before confirming an input.

5.3.4 Gaze Interaction. Due to WebXR limitations, we simulated gaze using head orientation to approximate the user's point of regard. UI elements are treated as collidable objects in 3D space, and selection is inferred based on sustained gaze dwell over a selectable target. A selection is confirmed if the user maintains gaze for at least 3.5 seconds, minimizing accidental activation while preserving low-effort interaction. Future research may leverage native OpenXR or Unity for gaze+pinch interaction [54].

6 Applications

Sensible Agent's core capability—dynamically adapting *what* proactive assistance to offer and *how* to interact—directly addresses the critical challenge of *interaction friction* that often hinders the practical adoption of proactive AR agents. By intelligently selecting the content and interaction modalities optimized for minimal user efforts and disruption, Sensible Agent enables compelling AR applications that were previously cumbersome or impractical:

Context-Adaptive Routine Support. Sensible Agent modifies its behavior based on learned routines, context, and user proficiency.

- **Beginner *vs.* Expert Use**: For a user's very first gym visit (Figure 6A), the system offers multi-choice for assistance, facilitating user exploration. However, for subsequent routine workouts, particularly in a noisy environment where voice input is difficult, it could adapt to allow a simple gaze gesture at an icon to start/stop a timer (Figure 6B), minimizing disruption and leveraging a modality suitable for the context.
- **Learned Preferences**: During a first-time restaurant visit (Figure 6C), the system might proactively offer dish recommendations. If the user dismisses this (*e.g.*, via a head shake) and explicitly requests menu translation, Sensible Agent can infer this preference. On subsequent visits to similar venues (Figure 6D), it can prioritize offering translation assistance proactively, adapting the content of its assistance based on interaction history.

Opportunistic Suggestion. Beyond adapting existing interactions, Sensible Agent can opportunistically introduce users to relevant but unexplored system capabilities. For example, during clothes shopping (Figure 6E), a user might primarily use the AR agent for price checks or list management. Sensible Agent can monitor user

[4]WebXR APIs: https://immersiveweb.dev
[5]three.js: https://threejs.org

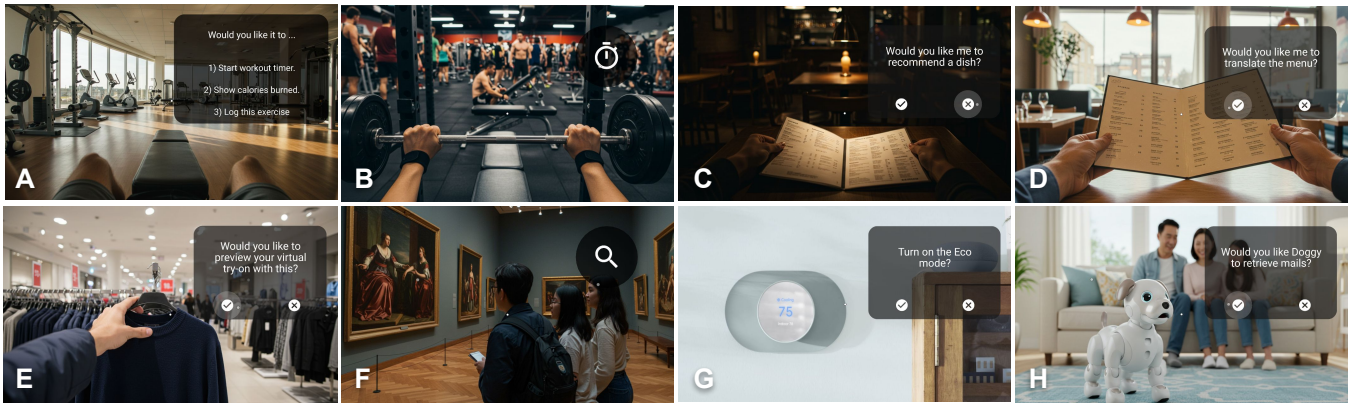

Figure 6: Applications. A-D): Sensible Agent's initial query (*I*) and repetitive query (*R*), based on the same daily scenarios. A) Gym visit-*I*. B) Gym visit-*R*. C) Restaurant order-*I*. D) Restaurant order-*R*. E) Novel feature suggestion: virtual try-on. F) Subtle cues: information retrieval. G) Effortless smart device control. H) Future application: Human-robot interaction.

activity and, during moments inferred as lower-urgency browse, proactively suggest a related but unused feature, such as virtual try-on. This facilitates feature discovery at moments when the user is likely receptive, without interrupting focused tasks.

Minimally Intrusive Augmentation. Sensible Agent framework preserves social flow by minimizing interference with the real-world procedures that they augment. In a museum setting (Figure 6F), users might be engaged in conversation with companions. Sensible Agent can provide access to supplementary information (*e.g.*, details about a painting via a subtle search icon) in a manner that requires minimal overt interaction, allowing users to access digital information without significantly disrupting the primary social activity.

Potential Extensions: Cross-Device Orchestration. As sensing capabilities improve and AR hardware becomes more integrated, we envision Sensible Agent acting as an orchestration engine. Future work could explore how the framework could dynamically select the most appropriate device and modality for a given task – potentially leveraging nearby surfaces as displays, integrating data from smart home sensors (Figure 6G), or coordinating actions with physical robotic agents (Figure 6H) – based on inferred user needs and context, further reducing interaction friction in complex, multi-device environments.

7 Preliminary Evaluation

We conducted a within-subjects preliminary evaluation to compare Sensible Agent against a baseline voice-controlled agent, modeled after existing systems like Project Astra [17]. This study aimed to surface insights into interaction efficiency, cognitive effort, and user preference across contextually varied scenarios.

We examined whether Sensible Agent would (1) reduce users' perceived cognitive load compared to explicit voice-based querying, (2) result in slower overall interaction time due to its two-step confirmation mechanism, and (3) be preferred in repetitive, context-varying situations where users experience situational impairments or fluctuating input/output availability. We also explored whether participants would default to familiar input modalities or adapt their interaction strategies based on context.

7.1 Participants

We recruited 10 participants (6 male, 4 female), aged 24–36 (μ = 30.6), from within our organization via internal mailing lists. All had prior experience using AR headsets (μ = 3.8) and moderate familiarity with voice-based assistants (μ = 3.6) on a 5-point Likert scale.

7.2 Apparatus and Experiment Design

The study was implemented on Project Moohan, an Android XR headset[6] using a WebXR-based prototype on Chrome v137. Participants experienced both AR (with video see-through) and VR environments and interacted with AI agent using gaze, voice, hand, and head gestures. A custom WebSocket-based control interface allowed the experimenter to manage the experimental flow remotely, including switching between AR and 360° video environments without requiring headset removal.

The experimental design overview is illustrated in Figure 7. Participants completed twelve scenarios across two system conditions: (1) Sensible Agent, which provided proactive assistance with unobtrusive, multimodal interactions; and (2) a *baseline* system that required users to initiate requests via voice commands, following a conventional assistant model (*e.g.*, *"What should I order?"*, *"Tell me about this exhibit."*). The system was Wizard-of-Oz controlled; participants tapped the headset to signal readiness, after which the experimenter triggered the appropriate system behavior (speech detection for baseline, environment detection for Sensible Agent).

Scenarios were divided into six high-level activities: three delivered as 360° videos (*e.g.*, reading a menu at a restaurant, shopping at a grocery store, commuting in bus), and three physically staged AR scenes (*e.g.*, cooking at a kitchen, working out at a gym, visiting the museum). Each activity was experienced in two forms: a *baseline unfamiliar version* and a *context variant* that modulated key context components that were explored in the workshop study and the data

[6] Android XR: https://www.android.com/xr

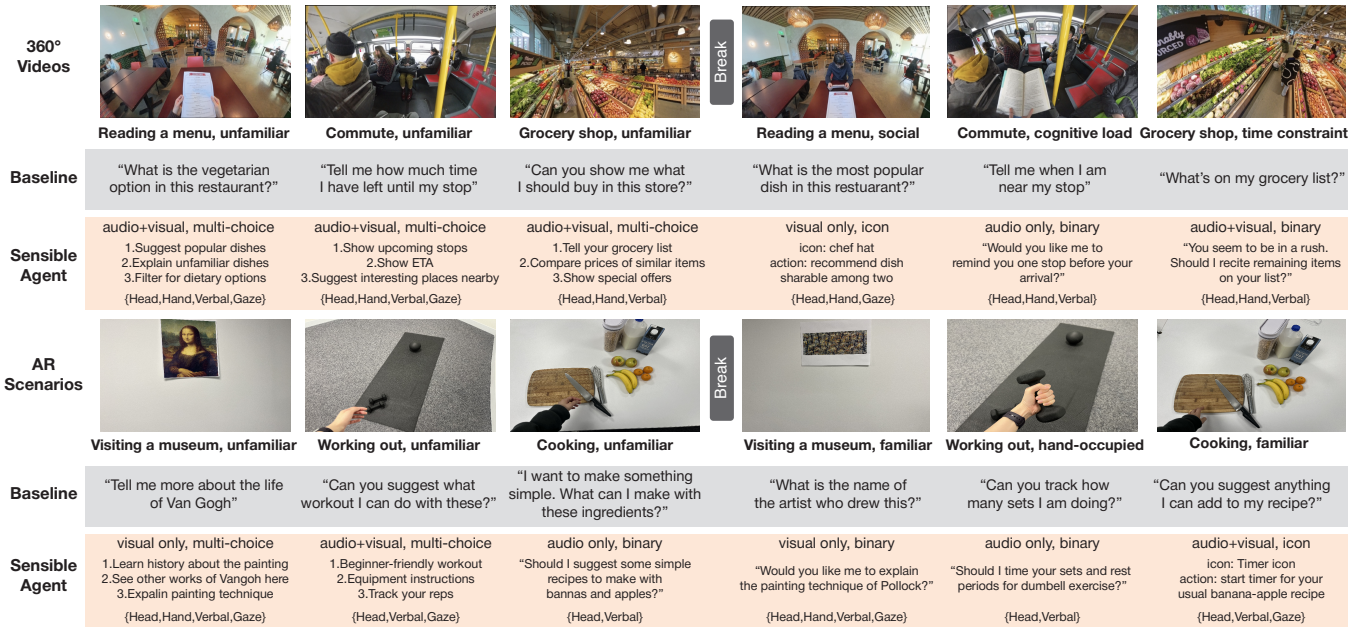

Figure 7: Experiment flow across six high-level activities. The top row shows screenshots from the three 360° video scenarios (reading a menu, commute, grocery shopping), and the bottom row depicts physically staged AR scenes (visiting museum, working out, cooking). Each scenario is labeled with its high-level activity and context variant (*e.g., unfamiliar, social setting*). Participants experienced all *unfamiliar* scenarios first, followed by their corresponding context variants, avoiding back-to-back repetition of the same activity to simulate naturalistic task switching. For each scenario, we include an example from the baseline condition (user-issued voice query) and a Sensible Agent response, showing the system-selected query type (icon, binary, or multi-choice), presentation modality (audio, visual, or audio+visual), a condensed version of the system-generated prompt, and the available input modalities based on context.

collection study, including social engagement, temporal urgency, or sensory constraints. For instance, a "reading a menu at a restaurant" scenario may be experienced first alone and then in a social setting. This paired design is to examine how interaction preferences shift when the same high-level activity occurs under altered contextual pressures.

The overall scenario flow was intentionally structured to avoid back-to-back repetitions of the same activity. As shown in Figure 7, participants first experienced three unfamiliar scenarios across different activities. After a short break, they encountered the corresponding three variant scenarios. This design aimed to minimize repetition effects and maintain participant engagement across tasks, while enabling direct within-subject comparison of behavior across unfamiliar and variant contexts.

7.3 Procedure

Each study session lasted approximately 45 minutes. Following a short tutorial on the interaction modalities supported in the system (*e.g.*, head gestures, hand gestures, gaze, or NLCS), participants completed six scenarios in total; three *unfamiliar* baseline versions followed by their paired *context variants*, each drawn from a different high-level activity (*e.g.*, cooking, commuting). To minimize learning or carryover effects, participants never encountered both versions of the same activity back-to-back. The scenario order was consistent across participants and is shown in Figure 7.

At the beginning of each trial, participants were briefly narrated the contextual framing of the scenario, including relevant conditions such as task familiarity, social engagement, or time pressure. During each scenario (lasting 2–3 minutes), participants responded to agent prompts using the available input modalities. In Sensible Agent conditions, the system adaptively enabled modalities based on context, and participants were free to respond using any that were available. Participants interacted with the agent until a system response was completed.

System conditions were counterbalanced across participants using a Latin square. After each system condition, participants completed the NASA-TLX [18] and System Usability Scale (SUS) [6] questionnaires. We also logged interaction time (from prompt trigger to system response) and recorded the modality used for each response.

At the end of the entire session, participants in a brief semi-structured interview. We asked about overall system preferences, perceived effort, comfort with interaction modalities, and expectations for proactive agent behavior in everyday contexts.

7.4 Results

We report both parametric (paired-sample t-test) and non-parametric (Wilcoxon signed-rank test) results, based on Shapiro-Wilk normality checks. All quantitative analyses are exploratory and not

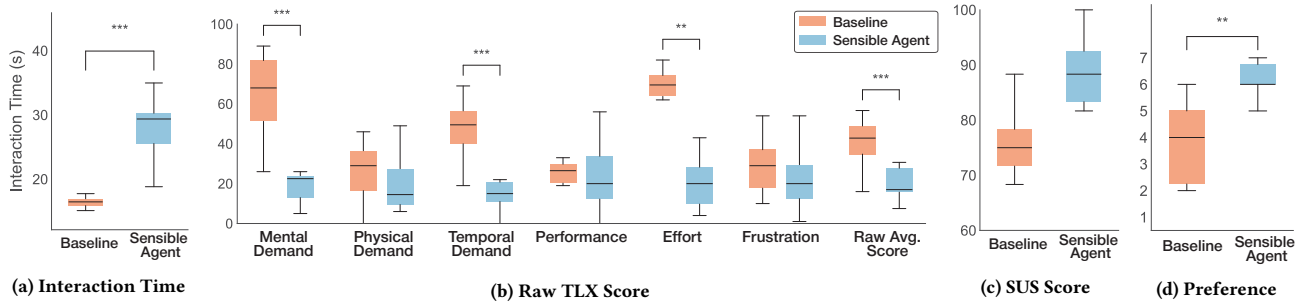

Figure 8: Quantitative analysis of (a) interaction time, (b) Raw TLX scores, (c) SUS scores, and (c) preference measures in our user study. The statistic significance is annotated with *, **, or *** (representing $p < .05$, $p < .01$, and $p < .001$, respectively).

corrected for multiple comparisons. Effect sizes (Cohen's d or rank-biserial r) are included for transparency (See Figure 8).

Interaction Time. Interaction was faster in the baseline voice-query condition (μ = 16.43s, σ = 0.84) than in the Sensible Agent condition (μ = 28.54s, σ = 4.85), $t(9) = -7.54$, $p < .001$, $d = -2.51$. This trend is expected due to Sensible Agent's two-step interaction flow, where the system first presents a suggested query based on context and the user then confirms or modifies it, in contrast to the baseline system where users immediately issue a voice command.

Cognitive Load (NASA-TLX). Sensible Agent showed lower mental demand (μ = 21.10, σ = 11.57) than the baseline (μ = 65.00, σ = 20.19), $t(9) = 6.40$, $p < .001$, $d = 2.03$. Participants also reported lower temporal demand with Sensible Agent (μ = 16.00, σ = 10.12) versus baseline (μ = 46.20, σ = 16.61), $t(9) = 5.43$, $p < .001$, $d = 1.72$.

Effort scores showed a similar trend: $W = 1.00$, $p = .0039$ (Wilcoxon), with Sensible Agent rated lower (μ = 20.30, σ = 12.79) than baseline (μ = 67.20, σ = 14.70). Participants frequently mentioned the ease of interaction as a key factor. Consistent differences were observed for physical demand ($p = .18$), performance satisfaction ($p = .65$), or frustration ($p = .23$).

Total Raw-TLX scores were lower for Sensible Agent (μ = 20.55, σ = 8.42) than baseline (μ = 43.27, σ = 9.73), $t(9) = 6.76$, $p < .001$, $d = 2.14$, reflecting a consistent pattern of reduced cognitive burden across conditions.

Usability (SUS). SUS scores were calculated using the standard procedure ($\sum$ item scores $\times$ 2.5, yielding a range of $[0, 100]$). There was no observed difference in SUS scores between the baseline (μ = 76.67, σ = 5.93) and Sensible Agent (μ = 81.33, σ = 6.58), $W = 11.00$, $p = .11$. Both conditions averaged above 71.4, which maps to a "Good" usability rating [4]. Detailed subscale scores are provided in Appendix C.

User Preferences. In a 7-point Likert scale, participants expressed a preference for Sensible Agent (μ = 6.00, σ = 0.94) over the baseline (μ = 3.80, σ = 1.48), $W = 0.00$, $p = .0074$. Seven of ten participants expressed that the proactive and unobtrusive nature of Sensible Agent made interaction more engaging.

Interaction Patterns. We observed notable patterns in how participants interacted with the multi-choice panel across different context scenarios. In the initial round of unfamiliar scenarios, those presented first in both the 360° and AR settings, participants selected different input modalities when prompted to confirm agent suggestions: 3 participants used voice, 2 used head gestures, and 4 used hand input. Six out of ten participants consistently used the same modality throughout this round, suggesting an early anchoring effect in input behavior. The remaining four participants changed modalities during the round, often out of exploratory intent.

Situational factors influenced modality choice. In scenarios where participants were physically holding objects—such as the gym (holding a dumbbell) or cooking (manipulating ingredients)—six participants switched to hands-free modalities like head or voice input, indicating sensitivity to situational input constraints (SSID). Two participants preferred head gestures, citing familiarity with similar gestures on consumer devices like Apple's AirPods. However, they noted that the tilting gesture required some adaptation. Gaze-based input was met with mixed responses; five participants reported that gaze sometimes conflicted with reading the panel content, leading them to look away intentionally to avoid accidental selections.

7.5 Qualitative Feedback

We summarize key takeaways from post-trial interviews, focusing on user perceptions of input, prompt format, and interactional fluency.

Preference for unobtrusive input. Five participants (P1, P2, P5, P8, P9) appreciated being able to respond using minimal-effort inputs such as head gestures or short verbal confirmations. P2 noted, *"I liked how I could reply in almost any way I wanted."* P9, *"It works well because the agent asked me something ... I didn't even have to think to ask. It was natural to just nod and keep going."* P1 highlighted the benefit of quick interactions: *"I could answer fast and move on. That's what makes it feel helpful instead of annoying."* P8 added, *"I loved how little effort I had to give to respond, which makes even more sense in situations where I have a friend who wants me to focus on them during a conversation."* Three participants (P3, P6, P10) expressed a desire for lightweight confirmation that their input had been successfully recognized.

Prompt format preferences. All ten participants found multi-choice prompts especially helpful in unfamiliar situations. P4 stated, *"The choices were helpful when I didn't know what I want. I could just pick."* Opinions on icon-only formats were mixed; while some appreciated their brevity, P7 mentioned interpretability challenges: *"It took me three seconds to figure out what the icon meant."*

Alignment with social interaction patterns. Several participants (P3, P8, P10) described the interaction style as resembling a casual conversation, where suggestions were context-aware and effort-free. As P10 put it, *"It felt like just talking to someone who already knows what I might want."*

8 Discussion

Our study reveals that proactive agents, when equipped with unobtrusive multimodal interfaces, not only reduce user effort but also reshape how users perceive and engage with digital assistants. We reflect on these broader implications and discuss how our framework can evolve based on observed behaviors from both studies.

8.1 Proactivity as a Social Cue

While prior work on proactive agents has focused on reducing friction or predicting user needs, our findings suggest that proactivity may also shape how users perceive the agent as a social presence. Seven out of ten participants reported that they found interactions with Sensible Agent to be more engaging or even enjoyable beyond simply requiring less effort.

The use of subtle, non-verbal input methods, such as nodding or tilting the head, further contributed to this perceived naturalness. Participants likened these gestures to the kinds of acknowledgments they use in everyday social interactions. These results point to a broader design opportunity: proactively adaptive systems may benefit from aligning more closely with human social cues, not only to reduce effort, but to foster rapport and interactional fluency.

8.2 Modality Fusion Based on Situational Weighting

While our current framework allows users to respond using multiple modalities, input signals are handled independently and without coordination. During our workshop study, two expert participants proposed that the system could go further by integrating multiple modalities using situational weighting. For instance, in real-world environments, users may emit overlapping or even conflicting cues—such as unintentionally nodding while verbally responding "*no*", or glancing away while affirming an option aloud. Such ambiguities are difficult to resolve without considering environmental context, task state, and historical user behavior.

This insight suggests an extension to our framework where each input modality is evaluated not only by its raw signal, but by its reliability under the current situational context. For example, voice input may be down-weighted in noisy environments, while hand gestures may be deprioritized when the user is physically constrained. This would allow the agent to compute a weighted confidence score across inputs, resulting in more robust intent inference. Building in such a multimodal arbitration layer—responsive to situational impairments and social context—would position proactive agents to operate more effectively in complex, dynamic environments.

In safety-critical or task-oriented contexts such as equipment repair or medical triage, such multimodal arbitration becomes essential—for both robustness and to modulate intrusiveness and timing. Our modular architecture could incorporate a task monitor that tracks procedural stages and interruptibility signals (*e.g.*, idle hands, pause in activity), enabling the agent to shift from ambient assistance to structured, stepwise support. This would allow proactive prompting to align not just with user availability, but with task flow.

8.3 Modeling Modality Preferences

While Sensible Agent currently treats each proactive interaction as a discrete event, several participants attempted to engage in follow-up utterances or gestures, suggesting a desire for extended, multi-turn exchanges. This highlights an opportunity to extend our framework by incorporating a temporal layer that tracks dialogue state and user responsiveness across turns.

One open design question is how agents should adapt presentation strategies during sustained interactions. Should the same modality persist across turns to support continuity and reduce surprise, or should the system adjust dynamically to maintain engagement or match user behavior? Extending our framework to include a temporal rhythm model could enable systems to better scaffold ongoing interactions—particularly in context-rich environments where attention and modality availability fluctuate.

In addition to managing temporal rhythm, agents must also account for longer-term user preferences that may not align with situationally optimal choices. For instance, a user might consistently prefer hand gestures over voice even in quiet environments. Our framework's notion of a *user action prior* (Figure 2 already supports situational preference modeling; future extensions could incorporate longer-term adaptation through techniques like reinforcement learning or implicit feedback tracking. This would enable the system to personalize modality strategies based not just on immediate context, but on evolving user tendencies and habits.

9 Limitations

Our current prototype and study were designed to validate core interaction principles; however, several areas remain open for future extension and evaluation. First, while our design supports personalization and contextual adaptation, our current prototype does not model user history or longitudinal preferences. Incorporating historical interaction patterns—such as preferred modalities, timing preferences, or context-specific behaviors—could enable more personalized and anticipatory agent behavior over time.

Second, we did not model the precise timing of proactive prompts within a given situation. Prior work in procedural task guidance has explored step-aware interventions where action boundaries are clearly defined [3, 33]. However, our target contexts involved open-ended activities (*e.g.*, browsing a museum, ordering food) where task state is fluid and interruption thresholds are less well-defined. Determining the optimal timing for intervention in such scenarios remains an open challenge and a promising extension to our framework.

Third, while our study compared to a conventional voice-query baseline, we did not include a condition where assistance was delivered through always-visible multi-choice interfaces using conventional XR interactions (*e.g.*, point + pinch) or microgestures (*e.g.*, gaze + thumb-to-finger swipe [53]). Comparing against such persistent UI paradigms could help isolate the specific contribution of unobtrusive, gaze- or head-based modalities to user experience and perceptions of agent presence. Additionally, our sample size

was limited (n=10), and the findings should be interpreted as preliminary and exploratory. While we observed consistent trends, future work should expand to assess generalizability and long-term effectiveness.

Finally, the scenarios used in our study focused on everyday, repeated contexts—such as grocery shopping or exercising—but did not include task-oriented settings with subsequent steps to take or high-stakes outcomes. Future work could explore how proactive, multimodal agents function in domains such as collaborative work, task guidance, or healthcare, where expectations and risks differ.

10 Conclusion and Future Work

In this paper, we address the critical challenge of interaction friction hindering the adoption of proactive AR agents. Existing approaches often rely on explicit, burdensome interactions unsuitable for many real-world contexts. Our work introduced Sensible Agent, a framework demonstrating the feasibility and benefit of dynamically adapting both the content ("what") and modality ("how") of proactive assistance to achieve unobtrusive interaction. By leveraging multimodal sensing and LMM reasoning, Sensible Agent selects contextually appropriate actions and interaction methods designed to minimize user effort. Our evaluation confirmed that this dynamic adaptation significantly reduces perceived intrusiveness and interaction burden compared to conventional methods, paving the way for more seamless human-AI collaboration in AR.

While Sensible Agent represents a significant step towards effortless proactive AR, several avenues warrant further exploration. Future work should focus on expanding the repertoire of unobtrusive interaction modalities beyond the current set (e.g., subtle haptics, ambient visualizations) and exploring adaptation within modalities (e.g., varying the level of detail). Developing rigorous benchmarks and standardized metrics to evaluate the effectiveness and unobtrusiveness of proactive AR agents remains a key challenge for the community. Furthermore, extending the framework to provide more personalized, long-term recommendations based on inferred user goals and preferences presents an exciting direction. Finally, integrating Sensible Agent's principles into broader ambient computing [50] and mirrored world [15] environments could unlock truly pervasive, yet respectful, proactive assistance.

Acknowledgments

We would like to thank Zhongyi Zhou, Vikas Bahirwani, Jessica Bo, Zheng Xu, Renhao Liu for their feedback and discussion on our early-stage proposal. We thank Alex Olwal for founding and directing the Interaction Lab at Google Research, directing the Augmented Language workstream, and pioneering research in [49] that inspired this work.

References

[1] Steven Abreu, Tiffany D Do, Karan Ahuja, Eric J Gonzalez, Lee Payne, Daniel McDuff, and Mar Gonzalez-Franco. 2024. Parse-Ego4d: Personal Action Recommendation Suggestions for Egocentric Videos. *ArXiv Preprint ArXiv:2407.09503* (2024). doi:10.48550/arXiv.2407.09503

[2] Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. 2022. Agreement-on-the-Line: Predicting the Performance of Neural Networks Under Distribution Shift. *Advances in Neural Information Processing Systems* 35 (2022), 19274–19289. doi:10.48550/arXiv.2206.13089

[3] Saptarashmi Bandyopadhyay, Vikas Bahirwani, Lavisha Aggarwal, Bhanu Guda, Lin Li, and Andrea Colaco. 2025. YETI (YET to Intervene) Proactive Interventions by Multimodal AI Agents in Augmented Reality Tasks. *ArXiv Preprint ArXiv:2501.09355* (2025). doi:10.48550/arXiv.2501.09355

[4] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *Journal of Usability Studies* 4, 3 (2009), 114–123.

[5] Keping Bi, Qingyao Ai, and W Bruce Croft. 2021. Asking Clarifying Questions Based on Negative Feedback in Conversational Search. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 157–166. doi:10.1145/3471158.3472232

[6] John Brooke et al. 1996. SUS-A Quick and Dirty Usability Scale. *Usability Evaluation in Industry* 189, 194 (1996), 4–7.

[7] Arthur Caetano, Alyssa Lawson, Yimeng Liu, and Misha Sra. 2023. ARLang: An Outdoor Augmented Reality Application for Portuguese Vocabulary Learning. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 1224–1235. doi:10.48550/arXiv.2411.05211

[8] Runze Cai, Nuwan Janaka, Hyeongcheol Kim, Yang Chen, Shengdong Zhao, Yun Huang, and David Hsu. 2025. AiGet: Transforming Everyday Moments Into Hidden Knowledge Discovery With AI Assistance on Smart Glasses. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–26. doi:10.48550/arXiv.2501.16240

[9] Michelle Carney, Barron Webster, Irene Alvarado, Kyle Phillips, Noura Howell, Jordan Griffith, Jonas Jongejan, Amit Pitaru, and Alexander Chen. 2020. Teachable Machine: Approachable Web-Based Tool for Exploring Machine Learning Classification. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM. doi:10.1145/3334480.3382839

[10] Limin Chen, Zhiwen Tang, and Grace Hui Yang. 2020. Balancing Reinforcement Learning Training Experiences in Interactive Information Retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1525–1528. doi:10.1145/3397271.3401200

[11] Eugene Cho. 2019. Hey Google, Can I Ask You Something in Private?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–9. doi:10.1145/3290605.3300488

[12] Bidyut Das, Mukta Majumder, Arif Ahmed Sekh, and Santanu Phadikar. 2022. Automatic Question Generation and Answer Assessment for Subjective Examination. *Cognitive Systems Research* 72 (2022), 14–22.

[13] Stephan Diederich, Alfred Benedikt Brendel, Stefan Morana, and Lutz Kolbe. 2022. On the Design of and Interaction With Conversational Agents: An Organizing and Assessing Review of Human-Computer Interaction Research. *Journal of the Association for Information Systems* 23, 1 (2022), 96–138. doi:10.1145/3671151.3671189

[14] Mustafa Doga Dogan, Eric J Gonzalez, Karan Ahuja, Ruofei Du, Andrea Colaço, Johnny Lee, Mar Gonzalez-Franco, and David Kim. 2024. Augmented Object Intelligence With XR-Objects. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–15. doi:10.1145/3654777.3676379

[15] Ruofei Du, David Li, and Amitabh Varshney. 2019. Geollery: A Mixed Reality Social Media Platform. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI, 685)*. ACM, 13 pages. doi:10.1145/3290605.3300915

[16] Ruofei Du, Alex Olwal, MathieuLe Goc, Shengzhi Wu, Danhang Tang, Yinda Zhang, Jun Zhang, DavidJoseph Tan, Federico Tombari, and David Kim. 2022. Opportunistic Interfaces for Augmented Reality: Transforming Everyday Objects Into Tangible 6DoF Interfaces Using Ad Hoc UI. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI, 183)*. ACM, 1–4. doi:10.1145/3491101.3519911

[17] Google DeepMind. 2024. Project Astra. https://deepmind.google/technologies/project-astra/. Accessed: 2025-04-09.

[18] SG Hart. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Human Mental Workload/Elsevier* (1988).

[19] Steven J Henderson and Steven Feiner. 2008. Opportunistic Controls: Leveraging Natural Affordances As Tangible User Interfaces for Augmented Reality. In *Proceedings of the 2008 ACM Symposium on Virtual Reality Software and Technology*. 211–218. doi:10.1145/1450579.1450625

[20] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *ArXiv Preprint ArXiv:1704.04861* (2017). doi:10.48550/arXiv.1704.04861

[21] Faria Huq, Zora Zhiruo Wang, Frank F Xu, Tianyue Ou, Shuyan Zhou, Jeffrey P Bigham, and Graham Neubig. 2025. COWPILOT: A Framework for Autonomous and Human-Agent Collaborative Web Navigation. *ArXiv Preprint ArXiv:2501.16609* (2025). doi:10.48550/arXiv.2501.16609

[22] Razan Jaber, Sabrina Zhong, Sanna Kuoppamäki, Aida Hosseini, Iona Gessinger, Duncan P Brumby, Benjamin R Cowan, and Donald McMillan. 2024. Cooking With Agents: Designing Context-Aware Voice Interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–13. doi:10.1145/3613904.3642183

[23] Arata Jingu, Yudai Tanaka, and Pedro Lopes. 2023. Lipio: Enabling Lips As Both Input and Output Surface. In *Proceedings of the 2023 CHI Conference on Human*

Factors in Computing Systems. 1–14. doi:10.1145/3544548.3580775

[24] Janne Kaeder, Maurizio Vergari, Verena Biener, Tanja Kojić, Jens Grubert, Sebastian Möller, and Jan-Niklas Voigt Antons. 2024. Working with Mixed Reality in Public: Effects of Virtual Display Layouts on Productivity, Feeling of Safety, and Social Acceptability. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 740–748. doi:10.1109/ISMAR62088.2024.00089

[25] Naoki Kimura, Tan Gemicioglu, Jonathan Womack, Richard Li, Yuhui Zhao, Abdelkareem Bedri, Zixiong Su, Alex Olwal, Jun Rekimoto, and Thad Starner. 2022. SilentSpeller: Towards Mobile, Hands-Free, Silent Speech Text Entry Using Electropalatography. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19. doi:10.1145/3491102.3502015

[26] Rick Kjeldsen. 2001. Head Gestures for Computer Control. In *Proceedings IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*. IEEE, 61–67. doi:10.1109/RATFG.2001.938911

[27] Rachna Konigari, Saurabh Ramola, Vijay Vardhan Alluri, and Manish Shrivastava. 2021. Topic Shift Detection for Mixed Initiative Response. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 161–166. doi:10.18653/v1/2021.sigdial-1.17

[28] Abhishek Kumar, Lik-Hang Lee, Jagmohan Chauhan, Xiang Su, Mohammad A Hoque, Susanna Pirttikangas, Sasu Tarkoma, and Pan Hui. 2022. PassWalk: Spatial Authentication Leveraging Lateral Shift and Gaze on Mobile Headsets. In *Proceedings of the 30th ACM International Conference on Multimedia*. 952–960. doi:10.1145/3503161.3548252

[29] Nallapaneni Manoj Kumar, P. Ranjith Krishna, Pavan Kumar Pagadala, and N. M. Saravana Kumar. 2018. Use of Smart Glasses in Education-A Study. In *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)*. 56–59. doi:10.1109/I-SMAC.2018.8653666

[30] Wallace S Lages and Doug A Bowman. 2019. Walking With Adaptive Augmented Reality Workspaces: Design and Usage Patterns. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 356–366. doi:10.1145/3301275.3302278

[31] Jaewook Lee, Andrew D Tjahjadi, Jiho Kim, Junpu Yu, Minji Park, Jiawen Zhang, Jon E Froehlich, Yapeng Tian, and Yuhang Zhao. 2024. CookAR: Affordance Augmentations in Wearable AR to Support Kitchen Tool Interactions for People With Low Vision. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–16. doi:10.48550/arXiv.2407.13515

[32] Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S Rodriguez, and Jon E Froehlich. 2024. GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20. doi:10.1145/3613904.3642230

[33] Chenyi Li, Guande Wu, Gromit Yeuk-Yin Chan, Dishita Gdi Turakhia, Sonia Castelo Quispe, Dong Li, Leslie Welch, Claudio Silva, and Jing Qian. 2025. Satori: Towards Proactive AR Assistant With Belief-Desire-Intention User Modeling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–24. doi:10.1145/3706598.3714188

[34] Jiahao Nick Li, Yan Xu, Tovi Grossman, Stephanie Santosa, and Michelle Li. 2024. OmniActions: Predicting Digital Actions in Response to Real-World Multimodal Sensory Inputs With LLMs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22. doi:10.1145/3613904.3642068

[35] Sugang Li, Ashwin Ashok, Yanyong Zhang, Chenren Xu, Janne Lindqvist, and Macro Gruteser. 2016. Whose Move Is It Anyway? Authenticating Smart Wearable Devices Using Unique Head Movement Patterns. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 1–9. doi:10.1109/PERCOM.2016.7456514

[36] Zhipeng Li, Christoph Gebhardt, Yves Inglin, Nicolas Steck, Paul Streli, and Christian Holz. 2024. SituationAdapt: Contextual UI Optimization in Mixed Reality With Situation Awareness via LLM Reasoning. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–13. doi:10.1145/3654777.3676470

[37] Zixuan Li, Lizi Liao, and Tat-Seng Chua. 2024. Learning to Ask Critical Questions for Assisting Product Search. *ArXiv Preprint ArXiv:2403.02754* (2024). doi:10.48550/arXiv.2403.02754

[38] Lizi Liao, Ryuichi Takanobu, Yunshan Ma, Xun Yang, Minlie Huang, and Tat-Seng Chua. 2020. Topic-Guided Conversational Recommender in Multiple Domains. *IEEE Transactions on Knowledge and Data Engineering* 34, 5 (2020), 2485–2496. doi:10.1109/TKDE.2020.3008563

[39] Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive Conversational Agents in the Post-ChatGPT World. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3452–3455. doi:10.1145/3539618.3594250

[40] Tianjian Liu, Hongzheng Zhao, Yuheng Liu, Xingbo Wang, and Zhenhui Peng. 2024. Compeer: A Generative Conversational Agent for Proactive Peer Support. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–22. doi:10.1145/3654777.3676430

[41] XingyuBruce Liu, JiahaoNick Li, David Kim, Xiang'Anthony' Chen, and Ruofei Du. 2024. Human I/O: Towards a Unified Approach to Detecting Situational Impairments. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 18 pages. doi:10.1145/3613904.3642065

[42] Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards Conversational Recommendation Over Multi-Type Dialogs. *ArXiv Preprint ArXiv:2005.03954* (2020). doi:10.48550/arXiv.2005.03954

[43] Feiyu Lu and Doug A Bowman. 2021. Evaluating the Potential of Glanceable Ar Interfaces for Authentic Everyday Uses. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 768–777. doi:10.1109/VR50410.2021.00104

[44] Allan MacLean, Richard M Young, Victoria ME Bellotti, and Thomas P Moran. 2020. Questions, Options, and Criteria: Elements of Design Space Analysis. In *Design Rationale*. CRC Press, 53–105. doi:10.1080/07370024.1991.9667168

[45] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. 2021. Accuracy on the Line: On the Strong Correlation Between Out-Of-Distribution and In-Distribution Generalization. In *International Conference on Machine Learning*. PMLR, PMLR, 7721–7735. doi:10.48550/arXiv.2504.00186

[46] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable Conversational Reasoning With Attention-Based Walks Over Knowledge Graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 845–854. doi:10.18653/v1/p19-1081

[47] Florian Müller, Martin Schmitz, Daniel Schmitt, Sebastian Günther, Markus Funk, and Max Mühlhäuser. 2020. Walk the Line: Leveraging Lateral Shifts of the Walking Path As an Input Modality for Head-Mounted Displays. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15. doi:10.1145/3313831.3376852

[48] Takuya Nakata, Masahide Nakamura, Sinan Chen, and Sachio Saiki. 2024. Needs Companion: A Novel Approach to Continuous User Needs Sensing Using Virtual Agents and Large Language Models. *Sensors* 24, 21 (2024), 6814. doi:10.3390/s24216814

[49] Alex Olwal. 2009. *Unobtrusive Augmentation of Physical Environments: Interaction Techniques, Spatial Displays and Ubiquitous Sensing*. Ph. D. Dissertation. KTH.

[50] Alex Olwal and Artem Dementyev. 2022. Hidden Interfaces for Ambient Computing: Enabling Interaction in Everyday Materials Through High-Brightness Visuals on Low-Cost Matrix Displays. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–20. doi:10.1145/3491102.3517674

[51] OpenAI. 2024. Text-to-Speech API. https://platform.openai.com/docs/guides/text-to-speech. https://platform.openai.com/docs/guides/text-to-speech.

[52] Leonardo Pavanatto, Verena Biener, Jennifer Chandran, Snehanjali Kalamkar, Feiyu Lu, John J Dudley, Jinghui Hu, G Nikki Ramirez-Saffy, Per Ola Kristensson, Alexander Giovannelli, et al. 2024. Working in Extended Reality in the Wild: Worker and Bystander Experiences of XR Virtual Displays in Real-World Settings. *ArXiv Preprint ArXiv:2408.10000* (2024). doi:10.48550/arXiv.2408.10000

[53] Siyou Pei, David Kim, Alex Olwal, Yang Zhang, and Ruofei Du. 2024. UI Mobility Control in XR: Switching UI Positionings Between Static, Dynamic, and Self Entities. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 12 pages. doi:10.1145/3613904.3642220

[54] Ken Pfeuffer, Benedikt Mayer, Diako Mardanbegi, and Hans Gellersen. 2017. Gaze + Pinch Interaction in Virtual Reality. In *SUI '17: Proceedings of the 5th Symposium on Spatial User Interaction*. 99–108. doi:10.1145/3131277.3132180

[55] Jun Rekimoto. 2022. DualVoice: A Speech Interaction Method Using Whisper-Voice As Commands. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–6. doi:10.1145/3491101.3519700

[56] Jun Rekimoto. 2023. WESPER: Zero-Shot and Realtime Whisper to Normal Voice Conversion for Whisper-Based Speech Interactions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–12. doi:10.1145/3544548.3580706

[57] Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. 2023. Robots That Ask for Help: Uncertainty Alignment for Large Language Model Planners. *ArXiv Preprint ArXiv:2307.01928* (2023). doi:10.48550/arXiv.2307.01928

[58] Xuhui Ren, Hongzhi Yin, Tong Chen, Hao Wang, Zi Huang, and Kai Zheng. 2021. Learning to Ask Appropriate Questions in Conversational Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 808–817. doi:10.1145/3404835.3462839

[59] William Seymour and Max Van Kleek. 2021. Exploring Interactions Between Trust, Anthropomorphism, and Relationship Development in Voice Assistants. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–16. doi:10.1145/3479515

[60] Daniel Smilkov, Nikhil Thorat, Yannick Assogba, Charles Nicholson, Nick Kreeger, Ping Yu, Shanqing Cai, Eric Nielsen, David Soegel, Stan Bileschi, Michael Terry, Ann Yuan, Kangyi Zhang, Sandeep Gupta, Sarah Sirajuddin, D Sculley, Rajat Monga, Greg Corrado, Fernanda Viegas, and Martin M Wattenberg. 2019. TensorFlow.js: Machine Learning For The Web and Beyond. 1 (2019), 309–321. doi:10.48550/arXiv.1901.05350

[61] Sruti Srinidhi, Edward Lu, and Anthony Rowe. 2024. XAIR: An XR Platform That Integrates Large Language Models With the Physical World. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 759–767. doi:10.1109/ISMAR62088.2024.00091

[62] Theresa Jean Tanenbaum, Nazely Hartoonian, and Jeffrey Bryan. 2020. " How Do I Make This Thing Smile?" an Inventory of Expressive Nonverbal Communication in Commercial Social Virtual Reality Platforms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13. doi:10.1145/3313831.3376606

[63] Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric P Xing, and Zhiting Hu. 2019. Target-Guided Open-Domain Conversation. *ArXiv Preprint ArXiv:1905.11553* (2019). doi:10.48550/arXiv.2209.09746

[64] Zhiwen Tang, Hrishikesh Kulkarni, and Grace Hui Yang. 2021. High-Quality Dialogue Diversification by Intermittent Short Extension Ensembles. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 1861–1872. doi:10.18653/v1/2021.findings-acl.163

[65] Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The Information Bottleneck Method. *ArXiv Preprint Physics/0004057* (2000). doi:10.48550/arXiv.2507.07621

[66] Ching-Yi Tsai, Ryan Yen, Daekun Kim, and Daniel Vogel. 2024. Gait Gestures: Examining Stride and Foot Strike Variation As an Input Method While Walking. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–16. doi:10.1145/3654777.3676342

[67] Christian David Vazquez, Afika Ayanda Nyati, Alexander Luh, Megan Fu, Takako Aikawa, and Pattie Maes. 2017. Serendipitous Language Learning in Mixed Reality. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2172–2179. doi:10.1145/3027063.3053098

[68] Marilyn Walker and Steve Whittaker. 1990. Mixed Initiative in Dialogue: An Investigation into Discourse Segmentation. (June 1990), 70–78. doi:10.3115/981823.981833

[69] Xinru Wang, Mengjie Yu, Hannah Nguyen, Michael Iuzzolino, Tianyi Wang, Peiqi Tang, Natasha Lynova, Co Tran, Ting Zhang, Naveen Sendhilnathan, et al. 2025. Less or More: Towards Glanceable Explanations for LLM Recommendations Using Ultra-Small Devices. *ArXiv Preprint ArXiv:2502.19410* (2025). doi:10.48550/arXiv.2502.19410

[70] Nigel Ward. 2006. Non-Lexical Conversational Sounds in American English. *Pragmatics & Cognition* 14, 1 (2006), 129–182. doi:10.1075/pc.14.1.08war

[71] Maheshya Weerasinghe, Verena Biener, Jens Grubert, Aaron Quigley, Alice Toniolo, Klen Čopič Pucihar, and Matjaž Kljun. 2022. Vocabulary: Learning Vocabulary in Ar Supported by Keyword Visualisations. *IEEE Transactions on Visualization and Computer Graphics* 28, 11 (2022), 3748–3758. doi:10.48550/arXiv.2207.00896

[72] Jacob O Wobbrock. 2019. Situationally Aware Mobile Devices for Overcoming Situational Impairments. In *Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems*. 1–18. doi:10.1145/3319499.3330292

[73] Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive Human-Machine Conversation With Explicit Conversation Goals. *ArXiv Preprint ArXiv:1906.05572* (2019). doi:10.48550/arXiv.1906.05572

[74] Chengyuan Xu, Radha Kumaran, Noah Stier, Kangyou Yu, and Tobias Höllerer. 2024. Multimodal 3D Fusion and In-Situ Learning for Spatially Aware AI. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 485–494. doi:10.1109/ISMAR62088.2024.00063

[75] Jun Xu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Conversational Graph Grounded Policy Learning for Open-Domain Conversation Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1835–1845. doi:10.18653/v1/2020.acl-main.166

[76] Xuhai Xu, Anna Yu, Tanya R Jonker, Kashyap Todi, Feiyu Lu, Xun Qian, João Marcelo Evangelista Belo, Tianyi Wang, Michelle Li, Aran Mun, et al. 2023. XAIR: A Framework of Explainable AI in Augmented Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–30. doi:10.1145/3544548.3581500

[77] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2024. Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces. *ArXiv Preprint ArXiv:2412.14171* (2024). doi:10.48550/arXiv.2412.14171

[78] Chenchen Ye, Lizi Liao, Fuli Feng, Wei Ji, and Tat-Seng Chua. 2022. Structured and Natural Responses Co-Generation for Conversational Search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 155–164. doi:10.1145/3477495.3532063

[79] Bolin Zhang, Zhiying Tu, Yangqin Jiang, Shufan He, Guoqing Chao, Dianhui Chu, and Xiaofei Xu. 2021. DGPF: A Dialogue Goal Planning Framework for Cognitive Service Conversational Bot. In *2021 IEEE International Conference on Web Services (ICWS)*. IEEE, 335–340. doi:10.1109/ICWS53863.2021.00051

[80] Qinshi Zhang, Ruoyu Wen, Zijian Ding, Latisha Besariani Hendra, and Ray LC. 2024. Can AI Prompt Humans? Multimodal Agents Prompt Players' Game Actions and Show Consequences to Raise Sustainability Awareness. *ArXiv Preprint ArXiv:2409.08486* (2024). doi:10.48550/arXiv.2409.08486

[81] Yun Zhou, Tao Xu, Bertrand David, and René Chalon. 2016. Interaction On-the-Go: A Fine-Grained Exploration on Wearable PROCAM Interfaces and Gestures in Mobile Situations. *Universal Access in the Information Society* 15 (2016), 643–657. doi:10.1007/s10209-015-0448-6

[82] Yutao Zhu, Jian-Yun Nie, Kun Zhou, Pan Du, Hao Jiang, and Zhicheng Dou. 2021. Proactive Retrieval-Based Chatbots Based on Relevant Knowledge and Goals. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2000–2004. doi:10.1145/3404835.3463011

[83] Jie Zou, Yifan Chen, and Evangelos Kanoulas. 2020. Towards Question-Based Recommender Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 881–890. doi:10.1145/3397271.3401180

[84] Jie Zou and Evangelos Kanoulas. 2019. Learning to Ask: Question-Based Sequential Bayesian Product Search. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 369–378. doi:10.1145/3357384.3357967

A Action and Context Categories

This section details on the analysis of the action and context categories derived from our formative data collection study (see main paper, Section 5.2). The context variants were pre-defined by the scenarios presented to participants, while the action categories emerged from a thematic analysis of the collected user responses.

A.1 Context Categories

The context categories were systematically designed into our study scenarios to elicit proactive AI behaviors under various situational constraints. Each scenario was grounded in a high-level activity: commuting on a bus, cooking in a kitchen, working out at a gym, reading a menu at a restaurant, browsing a museum, or grocery shopping. The specific contextual variants for these activities were:
Default: The baseline version of the high-level activity with no specific contextual impairment or modulation. For example, in the 'restaurant' scenario, the user is simply reading a menu.

Temporal Urgency: The user faces a time constraint or pressure. For example, the user has to order quickly at the restaurant before a movie starts.

Familiarity-Based: The user is familiar with the environment or task. For example, the user is reading a menu at a restaurant they visit often.

Unfamiliarity-Based: The user is new to the environment or task. For example, the user is at a new restaurant in town for the first time.

Cognitive Load: The user is mentally or physically occupied with a secondary task. For example, reading a book while commuting on the bus or holding dumbbells while working out (hands-occupied).

Crowded: The surrounding environment is noisy and populated with other people. For example, the restaurant is busy and filled with patrons.

Socially-Engaged: The user is directly interacting with another person during the primary activity. For example, talking to a friend while deciding what to order at the restaurant.

Divergent Setting (diff): The user performs the same high-level activity but in a different type of venue, which may alter their needs or interaction patterns. For example, ordering from a casual café instead of a formal restaurant.

Environmental Changes (env): The immediate physical environment is altered in a way that creates a sensory challenge. For example, the user is ordering from a restaurant with dim lighting or trying to read an outdoor menu while it is raining.

The Crowded and Socially-Engaged categories introduce social constraints that an agent must navigate. Also, certain scenarios within Cognitive Load (e.g., hands-occupied) and Environmental Changes (e.g., dim lighting) represent forms of Sensory and Situational Impairments/Disabilities (SSIDs), where the user's ability to interact with standard interfaces is temporarily limited.

A.2 Action Categories

From the 937 user-generated responses, we identified eight distinct categories of proactive actions that users desired from the agent. These categories are defined as follows:

Suggest: Proposes a set of options or a single recommendation to aid the user's decision-making process (e.g., "Recommend a popular dish").

Remind: Surfaces timely, contextually relevant information that the user may have forgotten or needs to be aware of (e.g., "Remind me to get off at my stop two stops beforehand").

Guide: Provides turn-by-turn or step-by-step instructions to help the user complete a process (e.g., "Show me the step-by-step instructions on how to make this recipe").

Summarize: Processes a larger body of information to generate a new, condensed version. (e.g., "Summarize the description of this painting for me.").

Automate: Executes a pre-defined, multi-step workflow, often chaining together multiple actions that would otherwise need to be done manually. (e.g., "Log today's workout.").

Visual Augmentation: Augmenting mixed-reality overlays directly onto the physical world to enhance the user's perception (e.g., "Highlight the gluten-free items on the menu").

Information Retrieval: Fetches and presents a single, discrete piece of existing data without altering it. (e.g., in a museum, "Tell me who the artist is for this art piece").

Take App Action: Executes a single, explicit command within a software application. It is a direct trigger for one function. (e.g., "Play my workout playlist on Spotify").

A.3 Context Variant, Query Type, Activity Distribution

To provide a comprehensive overview of our dataset (N=937), we analyzed the distribution of user responses from two primary perspectives.

First, Table **??** details the complete breakdown of desired proactive behaviors. It cross-references the eight action categories (e.g., Suggest, Guide) with the ten contextual variants (e.g., Temporal Urgency, Cognitive Load), illustrating which specific actions were requested most frequently in each situation.

Second, Figure 9 visualizes the distribution of response counts across the different contextual variants (rows) and query types (columns). The data shows that preferences for query formats shift based on the user's situation. For example, higher counts are observed for binary-choice and multiple-choice formats in variants such as Divergent Setting, Socially-Engaged, and Temporal Urgency. A notable peak is observed for the combination of Socially-Engaged contexts and icon-based queries, which accounted for 69 entries.

A.4 Usefulness Rating

As shown in Figure 10, the perceived usefulness of query responses varies by both activity context and query format.

B System Latency and Prompting Efficiency

To assess the responsiveness of our system, we measured the agent-side latency of the Sensible Agent pipeline—from the moment a new context is constructed to the moment a complete proactive prompt (including reasoning, action type, query format, and presentation modality) is returned from the language model. This evaluation focuses specifically on generation time and does not include user response latency or downstream input handling.

Action Category	Context Category	Count
Automate	Default	6
Automate	Cognitive Load	13
Automate	Familiarity-Based	46
Automate	Unfamiliarity-Based	10
Automate	Divergent Setting	11
Automate	Temporal Urgency	13
Guide	Default	11
Guide	Unfamiliarity-Based	34
Guide	Divergent Setting	9
Guide	Environmental Changes	17
Information Retrieval	Default	7
Information Retrieval	Unfamiliarity-Based	39
Information Retrieval	Divergent Setting	22
Information Retrieval	Environmental Changes	6
Information Retrieval	Temporal Urgency	18
Remind	Default	26
Remind	Cognitive Load	18
Remind	Familiarity-Based	21
Remind	Divergent Setting	21
Remind	Temporal Urgency	33
Suggest	Default	43
Suggest	Cognitive Load	31
Suggest	Familiarity-Based	41
Suggest	Unfamiliarity-Based	14
Suggest	Socially-Engaged	134
Suggest	Crowded	52
Suggest	Divergent Setting	23
Suggest	Temporal Urgency	30
Summarize	Default	4
Summarize	Cognitive Load	6
Summarize	Temporal Urgency	4
Take App Action	Default	24
Take App Action	Cognitive Load	12
Take App Action	Socially-Engaged	42
Take App Action	Divergent Setting	24
Take App Action	Temporal Urgency	13
Visual Augmentation	Default	3
Visual Augmentation	Crowded	47
Visual Augmentation	Environmental Changes	16
Visual Augmentation	Divergent Setting	5

Table 2: Distribution of desired proactive actions (N=937) across eight context categories and eight action categories

Figure 9: Distribution of response counts across different contextual variants (rows) and query types (columns).

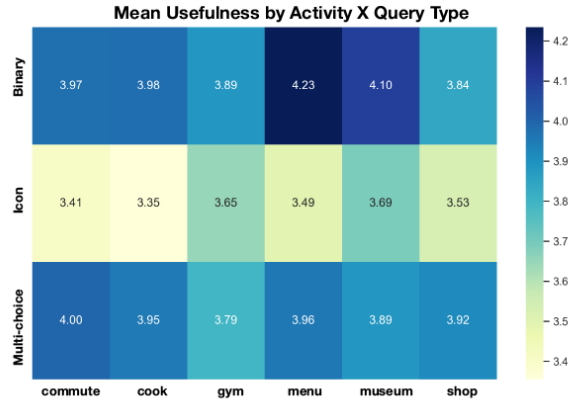

Figure 10: Heatmap showing the mean usefulness ratings of query responses across combinations of activity type (horizontal axis) and query type (vertical axis). Darker shades represent higher perceived usefulness, with binary queries during menu and museum activities receiving the highest ratings.

We conducted 20 trials using diverse test contexts spanning multiple high-level activities (e.g., grocery shopping, museum visit, working out) and context variants (e.g., time pressure, social setting, sensory impairment). Each trial used a prompt containing 6 few-shot exemplars selected based on contextual similarity heuristics, such as shared engagement type or environmental condition. The prompts were structured in a triplet format (context, reasoning, agent suggestion), and processed through GPT-4o via API.

Average generation latency was **6.2 seconds** ($\sigma = 0.8$) on a MacBook Pro (M1 Pro, 16GB RAM) with a stable internet connection. We observed minor variation based on the number and length of examples, but no critical delays for short-to-medium interactions (2–4 minutes) as used in our scenarios.

Compared to prior systems that employed structured LLM-based reasoning, our design reflects a trade-off between expressivity and responsiveness. *Human I/O* [41] reported an average latency of 19.95 seconds using GPT-4 and 7.33 seconds with GPT-3.5, using full Chain-of-Thought prompts focused on SIID detection. *OmniActions* [34] did not report exact latency, but leveraged fixed-size prompts for classification over a closed action label set. In contrast, our goal was to support open-context prompting with lightweight, interpretable few-shot examples while remaining within an acceptable delay for proactive interaction in everyday mobile settings.

Although our prompt set covers only a subset of the context-action space observed during data collection, we found that GPT-4o generalized well when exemplars shared key behavioral constraints. Future work may explore integrating retrieval-augmented generation or compact local models to further scale coverage while maintaining or improving latency.

C SUS Subscale Scores

We further examined each subscale component of the SUS questionnaire to investigate whether participants perceived differences between the two systems on specific usability aspects. While no subscale reached statistical significance after correction, we report descriptive comparisons to contextualize user preferences. For example, for the item *"I think that I would like to use this system frequently"*, ratings were slightly higher for the Sensible Agent condition ($\mu = 4.3$, $\sigma = 0.67$) compared to Baseline ($\mu = 2.1$, $\sigma = 1.2$), $W = 10.0$, $p = .09$. Similarly, for the item *"I found the system very cumbersome to use"*, Sensible Agent received higher ratings ($\mu = 3.3$, $\sigma = 0.48$) than Baseline ($\mu = 2.2$, $\sigma = 1.0$), $W = 12.0$, $p = .15$, suggesting a possible reduction in perceived complexity (Note that negatively worded items were reverse-scored prior to analysis.). No other items showed notable differences (all $p > .2$), including *"I felt very confident using the system"* and *"I would imagine that most people would learn to use this system very quickly"*, which were rated similarly across conditions.