SpeechLess: Micro-utterance with Personalized Spatial Memory-aware Assistant in Everyday Augmented Reality

Yoonsang Kim*
Stony Brook University

Devshree Jadeja†
Stony Brook University

Divyansh Pradhan‡
Stony Brook University

Yalong Yang§
Georgia Institute of Technology

Arie E. Kaufman¶
Stony Brook University

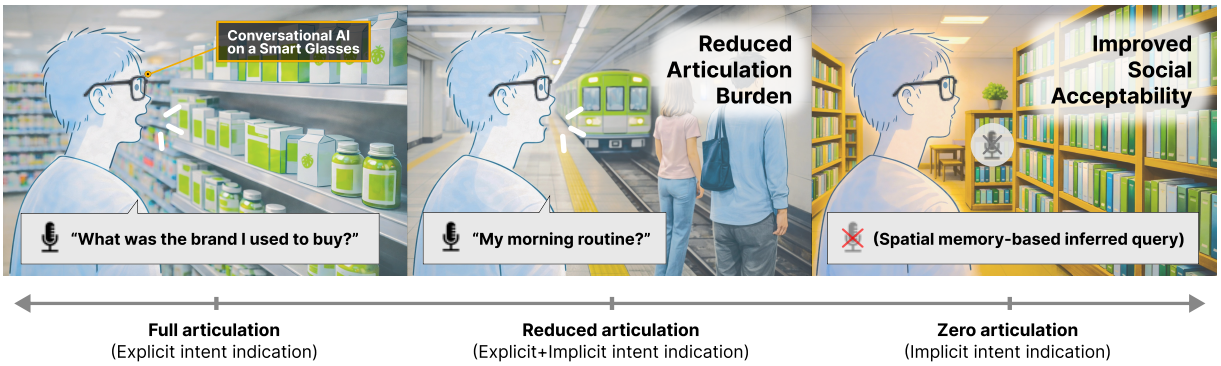

Figure 1: Concept illustration of SpeechLess. Users' queries are stored in the form of a "Spatial memory" along with their spatiotemporal, activity, and intent contexts. Later, under the similar/same routine context, these memories can be unobtrusively recalled with micro/zero-utterance, reducing the overhead of identical query repetition or minimal exposure of query contents to bystanders.

Abstract

Speaking aloud to a wearable AR assistant in public can be socially awkward, and re-articulating the same requests every day creates unnecessary effort. We present SpeechLess, a wearable AR assistant that introduces a speech-based intent granularity control paradigm grounded in personalized spatial memory. SpeechLess helps users "*speak less*," while still obtaining the information they need, and supports gradual explicitation of intent when more complex expression is required. SpeechLess binds prior interactions to multimodal personal context–space, time, activity, and referents–to form spatial memories, and leverages them to extrapolate missing intent dimensions from under-specified user queries. This enables users to dynamically adjust how explicitly they express their informational needs, from full-utterance to micro/zero-utterance interaction. We motivate our design through a week-long formative study using a commercial smart glasses platform, revealing discomfort with public voice use, frustration with repetitive speech, and hardware constraints. Building on these insights, we design SpeechLess, and evaluate it through controlled lab and in-the-wild studies. Our results indicate that regulated speech-based interaction, can improve everyday information access, reduce articulation effort, and support socially acceptable use without substantially degrading perceived usability or intent resolution accuracy across diverse everyday environments.

Index Terms: Wearables, Speech, Personal Assistant, Augmented Reality, Spatial Computing, Large Language Models, User Privacy.

*e-mail:yoonsakim@cs.stonybrook.edu
†e-mail:devshreehardik.jadeja@stonybrook.edu
‡e-mail:divyansh.pradhan@stonybrook.edu
§e-mail:yalong.yang@gatech.edu
¶e-mail:ari@cs.stonybrook.edu

1 Introduction

Augmented Reality (AR) provides a powerful mechanism for embedding digital information within the physical world, augmenting human perception and interaction [13]. By situating digital content in the user's surroundings, AR enables access to information that is inherently tied to physical context [38]. Recent advances in Large Language Models (LLMs) further extend this capability by supporting conversational interfaces that interpret natural language and respond in context. The convergence of AR and LLMs has opened new opportunities for intelligent assistants that not only situate content in the physical realm, but also reason about user intent through interaction, and enable more fluid and context-aware assistance [20,41,66]. The introduction of wearable AR extends this vision further, by embedding AI assistance into an always-available, first-person view form factor [57,72].

Wearable AR assistance can operate continuously within users' daily routines, where rich embodied context such as location, activity, and environment, naturally accompanies each moment of use. This continuous nature of wearable AR's context enables the system to better identify recurring everyday usage patterns and proactively anticipate the user's needs [39,43,48,54,59]. While existing wearable AR systems support proactive assistance, they largely provide limited mechanisms for users to regulate how their intent is specified or disambiguated when anticipatory inference is uncertain, socially inappropriate, or misaligned with moment-to-moment needs (e.g., binary confirmation, multiple-choice selection) [71]. To this end, wearable AR requires a new "*interaction paradigm*" that complements anticipatory inference, with a mechanism allowing context-rich expression of user intent–enabling users to dynamically specify or refine their needs while balancing convenience and reliability [18,27,62].

We introduce ***SpeechLess***, a wearable AR assistant that instantiates a new interaction paradigm for everyday use, while ensuring both convenience and robustness. We couple (1) long-term personalized spatial memory to represent recurring user activities, and (2) speech as a mechanism for dynamically adjusting the granularity of activity intent specification, depending on social context

and situational demands (e.g., Fig. 1–adjusting query style by social context). We design our proof-of-concept system informed by a 7-day formative study using a commercially available, conversational wearable AI assistant (Meta RayBan), which revealed recurring challenges with full speech-based interaction in public settings, frustration arising from repetitive verbal queries, and practical hardware limitations. To assess whether SpeechLess addresses these concerns, we evaluate the system across both public and private usage contexts, including crowded everyday environments and controlled laboratory settings. We examine how effectively SpeechLess supports intent resolution, reduces articulation burden, and maintains usability, under varying social and situational conditions.

By grounding information access in recurring everyday contexts, SpeechLess advances conversational assistance in wearable AR, toward a more context-aware and adaptive agent that allows users to regulate the granularity of intent expression (from micro to fully articulated verbal queries) based on social and situational constraints, supporting low-effort, socially acceptable interaction in daily activities. Our core contributions are as follows:

- **Interaction paradigm for everyday wearable AR:** The proposal of an interaction paradigm that allows users to control query intent granularity through speech, adapting to their needs, social, and situational context.
- **Personal spatial memory representation design:** A spatially grounded memory representation binding digital information with personal multimodal context–speech, space, time, actions, and referents–in the real-world for context-aware intent resolution.
- **Practical challenges on current smart wearables:** Empirical findings from a longitudinal in-the-wild study (with Meta RayBan), highlighting the current limitations of a speech-based smart wearable device in public/private spaces.
- **Demonstration and validation:** Validation of improved convenience and social acceptability, while maintaining accurate intent resolution under controlled visualization and socially constrained conditions (in-lab, crowds).

2 Related Work

AR at the Corner of Everyday Life. The growing integration of AR into daily life is reflected in popular applications such as Pokemon Go and Google Maps Live View [63, 69]. Prior works have explored how AR can support everyday use by situating information directly within physical environments. XR-Objects [20] extends the role of AR in everyday environments by retrieving knowledge with mobile AR via LLM. Others explore how AR enhances daily activities, including human memory recall and daily task support [5, 9, 23, 29, 33, 66]. These works show the potential for AR to be an integral part of our everyday life. Building on the notion of "everyday AR", we focus on capturing daily routine AR interactions to model long-term personal interests and their context.

Context-aware AI Assistant and LLM. LLMs are employed to contextualize multimodal input into conversational responses, enabling interpretation of intent and task-guidance [68, 76]. Prior works leverage LLM for accessibility or context-aware question answering [14, 30], or adopt LLM for proactive agents that anticipate user needs [28, 43, 74]. Sensible Agent shows how agent proactivity (system-triggered) and context-aware modality switching, can be used for enhancing socially acceptable AI [39]. Personalized contextual responses through user embeddings or situational cues, are also explored [3, 24, 32, 49, 56, 70]. These show how LLMs can ground multimodal inputs with context-aware responses. We leverage LLM to fuse user contexts (spatiotemporal, activity, and object) to infer intentions of user-spoken queries (user-initiated) [16, 37].

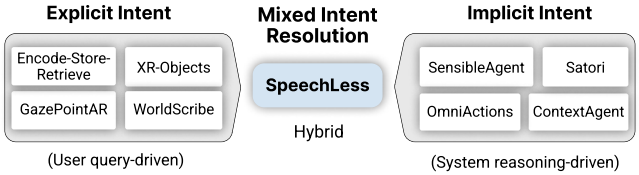

Figure 2: SpeechLess comprehends user intents using the benefits of both user-driven (explicit task prompting) and system-driven (proactive intent inference), intent resolution mechanisms.

Binding Information with Contextual Cues. Binding digital information to contextual cues–conversational, spatial, or visual–has been a longstanding goal in ubiquitous computing [45, 58]. OmniActions predicts user intents based on multi-source pattern learning [44]. Memoro demonstrates how a lightweight, audio-based assistant can log everyday conversations and support remembrance of prior queries through conversational replay [77]. Beyond conversational traces, prior works show the binding of AR content to physical contexts [4, 12, 20, 34, 47]. Encode-Store-Retrieve illustrates how egocentric perception can be encoded into language for episodic memory augmentation [66] and psychological studies of context-dependent memory highlight how environmental cues support recall [21, 65]. Other works leverage contextual and prior knowledge to support user assistance [1, 73]. Our work extends this trajectory by using spatially anchored personal memory not only for a reminder mechanism [29, 59], but also for filling missing implicit intent gaps in a spoken query (partially complete query), and leveraging it for convenience and unobtrusiveness in daily settings.

Use of Wearable AR in Public. Deploying an AI assistant on a wearable XR device such as smart glasses or an HMD can raise challenges of social comfort and privacy [6, 7, 39, 61]. Prior studies report that speech-based interaction in public can feel awkward, intrusive, or socially stigmatizing, while embodied gestural input may be impractical during mobility, or in public contexts [15, 51, 62, 69]. Research on privacy-preserving AR highlights how bystander exposure and unintended disclosure constrain real-world deployments [11, 17, 35, 36]. Glanceable AR emphasizes the need for lightweight, subtle, and socially unobtrusive modalities that reduce interaction effort and visible exposure [50]. Building on these insights, our work aims to design an interaction mechanism that minimizes audible articulation, improving social acceptability and protecting privacy during everyday public routines [22, 52].

As illustrated in Fig. 2, SpeechLess takes a *hybrid approach* to intent resolution, combining *user-driven explicit* intent specification (e.g., XR-Objects [20], GazePointAR [41]) and *system-driven implicit* intent extrapolation (e.g., Sensible Agent [39], OmniActions [44]). Users dynamically determine how much intent detail to convey through an expressive and natural interaction mode–***speech*** [62], while the system proactively extrapolates missing intent dimensions based on the user's personalized spatial context.

3 Design Rationale

We outline the motivation underlying our design, grounded in the need for a revised interaction paradigm for everyday wearable AR that leverages contextual intelligence to operate under social and situational constraints. We share the insights of our longitudinal formative study (Sec. 3.1), collected with a state-of-the-art wearable AI device (Meta RayBan). Then, we introduce the principles that guided the design of SpeechLess (Sec. 3.2).

3.1 Identifying the Practical Challenges of Wearable AI

To investigate the practical challenges of current wearable AI assistants, before extending them to AR, we conducted a seven-day formative study (N=8) using a commercially available wearable AI, Meta RayBan glasses. The aim behind the multi-day study was to examine how the devices fit into the participants in realistic daily

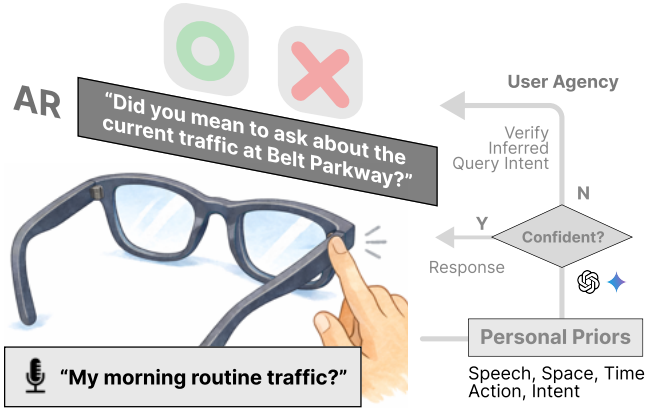

Figure 3: Conceptual illustration of device-to-user interface mapping. A minimalistic, hardware-based interface is preferred for the wearable; Human-in-the-loop for a transparent, trustworthy response.

routines across private and public environments. We focused on three research questions: **(Q1)** *What unique patterns does wearable smart glasses exhibit, compared to smartphones*; **(Q2)** *How do individuals perceive and manage themselves in public spaces while wearing the device* (when such form factor is not fully established in people's perception); and **(Q3)** *How do users feel about speech-based communication with the wearable AI assistant.*

Participants were recruited from diverse pool of professional backgrounds–researchers, software engineers, and students, and were instructed to wear the device for at least six hours a day (not consecutive), in private and public environments. They were guided to log their daily experiences, and provide qualitative feedback via semi-structured interviews, upon the completion of their participation. Feedback were analyzed using an inductive thematic analysis approach. Three co-authors independently reviewed, and grouped recurring challenges into high-level themes through discussion. Below, we share the coded themes:

Constraints from Social Norms. Verbal interaction was perceived as socially inappropriate in certain spaces such as libraries, classrooms (during lecture), or office settings. Participants avoided speaking aloud even when they needed the assistance (N=9). One participant described the dilemma: "*I wanted to look something up during class, but I just couldn't without disturbing everyone*". Hesitation due to workplace norms was reported also (N=3). These reflect a strong influence of situational norms on participants willingness to engage with the device.

Frustration in Articulation Burden. Participants expressed frustration with the repetitive invocation of the wake word ("Hey Meta"), describing it as cumbersome (N=6). Several resorted to hardware-based interaction (a physical tap on the side of the glasses) instead. Furthermore, participants reported cases where they were reluctant to use speech out loud for every question (N=5). These reveal the need for a more efficient mechanism to minimize the query effort, and for robustness in assistance triggering.

Limited Diverse Use Cases in Daily Routine. Participants reported that their daily routines left little room for novel interactions with the assistant. "*Didn't have anything new to ask everyday*". Many described repetitive schedules involving work-to-home commutes, where they found limited opportunities to ask new meaningful questions beyond their routine (e.g., weather, traffic, bus schedules) (N=6). Several participants described a lack of use cases when the same task could be achieved with their smartphone (N=2).

Social Awkwardness and Privacy Concerns. Participants described heightened public attention while using the glasses in public. Blinking lights intended as a visual indicator for assistants' listening mode, paradoxically drew more attention. "*People kept staring when the light blinked, even when I was not recording them, or at least I felt that way*". Privacy-related concerns were reported where bystanders misjudged them for recording (N=5), or assumed they were seeking human assistance when users spoke aloud alone (N=1). Other participants noted that the uncommon form factor (compared to mobile devices) and verbal queries, drew more attention than desired, making them hesitant to initiate assistance (N=6).

Hardware-induced Constraints. The physical and system constraints of the wearable device were pointed out. Short battery life–lasting less than 30 minutes during consecutive video recording (N=1), and paired smartphone overheating during data synchronization (N=2). Others brought up the limitation of recordings resolution, connection reliability, and form factor inconvenience (N=4).

3.2 Design Goals and Considerations

Our longitudinal in-the-wild formative study (Sec. 3.1) revealed several frictions in the everyday use of a wearable assistant. In the following section, we detail how each challenge motivated our design:

D1 *Socially Appropriate Intent Expression.* Fully articulated speech was found to limit the AI assistance in socially constrained environments, even when the assistance was desired. To address this, SpeechLess is designed to decouple assistance from mandatory verbal articulation. Speech is treated as a flexible mechanism for expressing a user's needs (or intent), allowing the user to vary how explicitly they communicate (from full to micro/zero-utterance), depending on social context or situational convenience. We aim to mitigate the discomfort of public speech while preserving access to expressive interaction when needed.

D2 *Spatially Grounded Personal Memories for Routines.* The informational needs of a wearable assistant were largely routine and repetitive, yet required users to repeatedly articulate the queries. We collect the user's personal context across everyday places and activities, identifying patterns in a long-term view. By binding user queries to personal multimodal context–space, time, action, intent, and referents–over time, we identify the plausible intent behind every user activity context [31, 34, 40].

D3 *Proactivity with User Controlled Disambiguation.* While our spatial memories can proactively predict the intent of an action without explicit user specification, the robustness and transparency are key components to ensure a satisfactory AR assistance [71]. Thus, we employ a human-in-the-loop, surfacing a confirmation interface (Yes/No) with rationale upon uncertainties (Fig. 3).

D4 *Minimal Interaction under Practical Constraints.* At the time of the study, the major hardware constraints of a wearable assistive device (Meta RayBan Gen-1), were battery drain (lasting 30 mins under consecutive video recording), overheating, and the social cost of verbal wake-word invocation ("*Hey Meta*"). Therefore, we favor a minimal, on-demand interaction model that reduces continuous sensing and computation load–no SLAM or gesture recognizer necessitating continuous frame access. As a minimalistic fallback mechanism to speech-based interaction, we map four button-based triggers mirroring the Meta RayBan tap gestures (single, double, triple, and tap-and-hold) (Fig. 3).

4 Implementation of SpeechLess

Interaction Paradigm: Speech as Granularity Control. Speech is treated as the primary channel for explicit intent expression due to its ability to convey user needs and context within a single interaction. It is an effective mode of communication to express intentions, especially for limited or no-screen devices such as smart AR wearables [62]. In SpeechLess, we leverage speech as a granularity control mechanism that provides users with direct agency

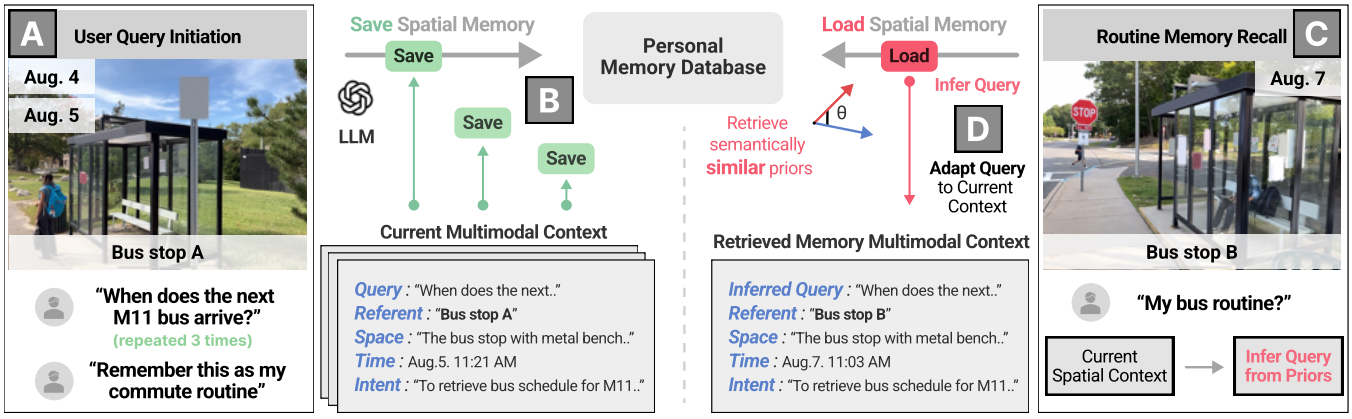

Figure 4: SpeechLess pipeline overview: Repeated queries **(A)** labeled as "commute routine" are stored in a personal memory data-storage **(B)**, where a memory recall request **(C)** retrieves similar contextual memories, and uses them to infer the intention behind a user's new query, and proactively adapt to the user's current context **(D)**. The context of bus schedule and personal routine are extrapolated from memories

over how explicitly their informational needs are specified. Users can selectively structure verbal queries to indicate intent, ranging from fully articulated queries with details, to partial core keyword-only utterances, and to zero-utterance interaction, depending on their situational context (D1). To support this flexibility, we combine the general knowledge reasoning from LLM agents with long-term, personalized spatial memory (space, time, activity, referent), to extrapolate missing information/gaps in user intent specification (D2). This design enables intent resolution even when verbal articulation is minimal or absent (Fig. 4). When agentic inference is uncertain, user feedback interface helps refine and confirm intent (D3). We mirror the interaction mechanisms of commercially available smart wearable glasses (Meta RayBan), including verbal queries, memory remembrance, and minimal hardware-based gesture mappings (D4), to assess the validity of our approach under practical conditions. We define three primary modes of our speech-based intent expression: *Full*, *Partial*, and *Zero* Utterance:

- **Full Utterance :** "When does the next M11 bus arrive?"
- **Partial Utterance :** "bus?" or "M11 bus?"
- **Zero Utterance :** (No speech–Reliant on prior memory)

Added intent cues

Figure 5: Varying granularity of speech inputs for intent expression.

Micro-utterance interaction (*Partial*, *Zero*) extrapolates missing intent dimensions using the user's priors–spatial memory. Users can add cues (at their discretion) for more explicit intent expression (Fig. 5) (**even within *Partial Utterance***: "bus" to "M11 bus")

System overview. SpeechLess consists of six modules that jointly deliver context-aware intent resolution through speech granularity control and personal spatial memory: (1) Query Decoding Agent interprets the user's interaction state by classifying the level of intent articulation and its type; (2) Contextual Dimension Encoder constructs a structured "dimension sketch" that represents the user's ongoing spatial activity, combining verbal cues with embodied context; (3) Personal Memory Store Manager incrementally builds long-term, personalized memories by binding prior interactions, inferred intents, and responses to spatiotemporal and referential contexts; (4) Memory Retriever loads relevant past episodes with contextual similarity to the user's current state, enabling reuse and intent extrapolation; (5) Answer Composer synthesizes concise responses by integrating the current scene with retrieved personal memories, and (6) User Verification Interface provides a light confirmation interface for uncertain intents. Note that the primary input channel of SpeechLess is speech supplemented by buttons as a fall-back mechanism to initiate a query or provide confirmations.

4.1 Query Decoding Agent

Query Decoding Agent serves as the entry point of the pipeline, ensuring that each incoming utterance (or zero-utterance) is routed appropriately. Every query is categorized into three types: *Question Answering*, *Remembrance*, and *Removal*, decomposing the request into a set of sub-tasks. This guarantees reasoning transparency in the pipeline, and deterministic control over each step (over an autonomous decision making agentic design).

Question Answering: represents a question-answering type. Within this category, we define three subtypes based on the level of verbal specification: *'Full Utterance'*–a query that explicitly states the user intent, and can be answered directly from the current view or prior knowledge/memory; *'Partial Utterance'*–under-specified or includes an implicit intent revealed only partially in the query, requiring intent inference drawn from prior user interactions; *'Zero Utterance'*–zero-verbal cue, where the system fully infers the user's intent from spatial (visuo-location) memory contexts.

Remembrance: applies when a user expresses intent in saving a memory ("*Can you note that..*" or "*Remember that..*") without asking for a response (unidirectional command). The scene description, object-of-focus, time, and intent are encoded into an episodic entry and prepared for storage directly via Memory Store Manager.

Removal: provides users the freedom to manage their saved memories. Upon a verbal request ("*Can you remove the memory on..*"), the system identifies the best-matching memory (semantically and keyword-based), summarizes it (for explainability), and requests user confirmation. Once confirmed, the memory is removed.

4.2 Contextual Dimension Encoder

Dimension Encoder extracts the user's multimodal input into a compact "dimension sketch." This sketch captures the user's current *space label*, *scene description*, *referent/object-of-focus*, *temporal moment*, *action intent*, and the *raw speech transcription*. These dimensions are bound jointly to form an instance of a "spatial memory." For under-specified intents under Question Answering queries (*Partial* or *Zero*), the encoder passes these dimensions to the intent resolution step, where it attempts to refine and infer the intent of the query based on the user's stored/prior spatial memories.

4.3 Memory Retriever

Once a speech input is classified as Question Answering, and requires prior context of the user's interaction (*Partial*, *Zero*), the module computes contextual similarity between the current dimension sketch and stored episodes. It surfaces up to five relevant final candidates ($k = 5$) to the user's intent and spatiotemporal context. The Retriever operates along: GPS proximity, semantic similarity

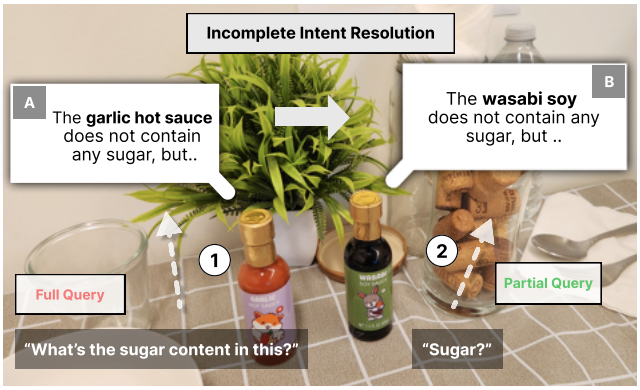

Figure 6: Proactive Intent Revision. SpeechLess can adapt a query using spatial memories. A query asking the sugar content of **(A)** can proactively be adapted for **(B)** upon a shift in the referent-of-interest.

(text, scene, and referent), and temporal recency/filter (if specified, e.g., time=="Tuesdays"); based on their semantic embeddings, re-ranking, and dimension fusion via Reciprocal Rank Fusion.

Static Source Prior Knowledge: If a query involves simple recalling of a prior 'as-is' memory, SpeechLess recalls knowledge directly from Memory Store Manager, combining retrieved episodes with the current scene. No external knowledge access (Web) is required, and responses are grounded in user's personal memories.

Live Source Prior Knowledge: If the user refers to a memory that involves a live, up-to-date content (e.g., weather, transit), an updated response is needed. The Retriever proactively updates the response of the retrieved memory via the Google Custom Search API [26], before returning it as candidate memories.

Fresh Knowledge (No Prior): If a requested information in the user's query does not have any prior relevant memories (or has low confidence), but was requested with *Partial* or *Zero Utterance*, which requires query intent inference, the generic knowledge of the LLM is used to infer the intent. Then, it is validated by user via User Verification Interface, for a trustworthy response.

4.4 Response Composer

The Response Composer Agent follows a RAG-like pipeline [42]. It aggregates five candidate memories received from Retriever, and together with the user query and current view, it generates an informed response. The agent is instructed to generate an adaptive response with respect to the user's current context (Fig. 6), with a concise answer (up to 30 words) to maintain simplicity for both visual and audio assistance. The composed answer is validated by an Answer Validation Agent that evaluates the response with a 1-10 confidence score. This response is transmitted to the client with a short rationale appended, following the principles of an explainable framework [71], ensuring that every reasoning step is accompanied by a rationale, guiding the user's understanding of "*why here*" and "*why now*". For a low confidence response, the user is asked to verify via Verification Interface.

4.5 Personal Memory Store Manager

Personal Memory Store Manager serves as a long-term, per-user episodic memory layer. Each entry binds the query, its response, and the dimension sketch that contextualizes the interaction (*scene description*, *object*, *time*, *intent*, and *spatial metadata*), in the form of a "Spatial memory." This storage grows as part of the user's routines. Its design privileges routine-specific recalls, ensuring that personal memories can be re-accessed in future contexts. Furthermore, the store operation is gated by User Verification, with deletions supported at any time, upon confirmation. A low-confidence response triggers a user correction/verification rather than being silently recorded, to maintain high-quality memory storage. A memory is stored by either Remembrance or Question Answering.

4.6 User Verification Interface and Visualization

This ensures that the intention inference remains accountable and user-centric. Each low-confidence response is paired with a binary confirmation–accept or reject (Yes/No), along with its rationale (Fig. 3). This human-in-the-loop mechanism allows users to curate their memory corpus with simplicity.

For visualization, query responses are presented either as screen-space UI (transcription of a verbal query) or as AR-anchored overlays (response). Users are given the option to toggle between the two modes (UI overlay vs AR situated text) allowing users to closely view distant AR content, while providing the option for a situated visual response. For the latter, we leverage LLM-based referent localization [25] to support open-vocabulary object detection. The identified referent is localized by combining visual and depth cues, enabling 3D placement of the AR content. This anchoring mechanism can act as a visual feedback that disambiguates the referent of the response, particularly when the user's query contains ambiguous references (under-specified verbal descriptions), where it is unclear which referent is associated with the response.

5 Use Cases and Applications

We illustrate how SpeechLess is used, to reduce social awkwardness, maintain continuity of routines, and enable minimal or silent interaction during everyday activities. All visuals in Fig. 7 were generated from in-app runs of SpeechLess, denoted as A1 to A6.

5.1 Home Routines: Reduced Efforts on Routine Tasks

Household activities involve recurrent tasks that benefit from memory continuity. As shown in A1, a note "*Remind me to water the plant on Tuesdays*.", can be simplified as "*plant?*." on the following Tuesday, while looking at the plant. SpeechLess retrieves prior episodes associated with its temporal dimension, "*Tuesday*" and its spatial (scene, referent, location) cues, producing a concise nudge: "*Plant watering on Tuesdays*." This shows how repetitive/routine queries can reduce users' verbal and remembrance effort.

5.2 Commute: Minimal Verbal Query among Crowd

Commuting is a recurring routine where users require timely updates on transportation schedules or traffic conditions. At a frequently visited bus stop (A2), the user can obtain an up-to-date bus schedule without speaking aloud in public. Instead of fully articulating, "*When does the next M11 bus arrive?*," "*M11 bus?*" or "*routine bus schedule*" (another memory-instance knowledge – Fig. 4-A)) is sufficient. SpeechLess proactively updates the live memory upon recall, ensuring that an accurate bus schedule is retrieved with minimal verbal cost in a socially sensitive context.

5.3 Classroom: Low-Profile Capture and Recall

Speaking aloud in a classroom or a library is socially inappropriate. SpeechLess enables low-profile interactions by supporting minimal utterances that are bound to the surrounding context. During a lecture, a student signals a "Remembrance" command while looking at course material, with a brief verbal note "Assignment". SpeechLess records the visual description of the user's view, timestamp, and the referent (a chapter in a book) as part of the personal memory. Later, upon looking at the same course material, SpeechLess can recall the assignment note anchored to the book (A3).

5.4 Grocery Shopping: Cross-space Information Recall

Grocery lists frequently involve updating items, set elsewhere outside the store (A4). At home, a user may say, *Remember to buy unsweetened soy milk from Walmart*." Later, at the store, when the user initiates a query: "*What did I say about buying something from Walmart?*" Despite the spatial context disparity between home and

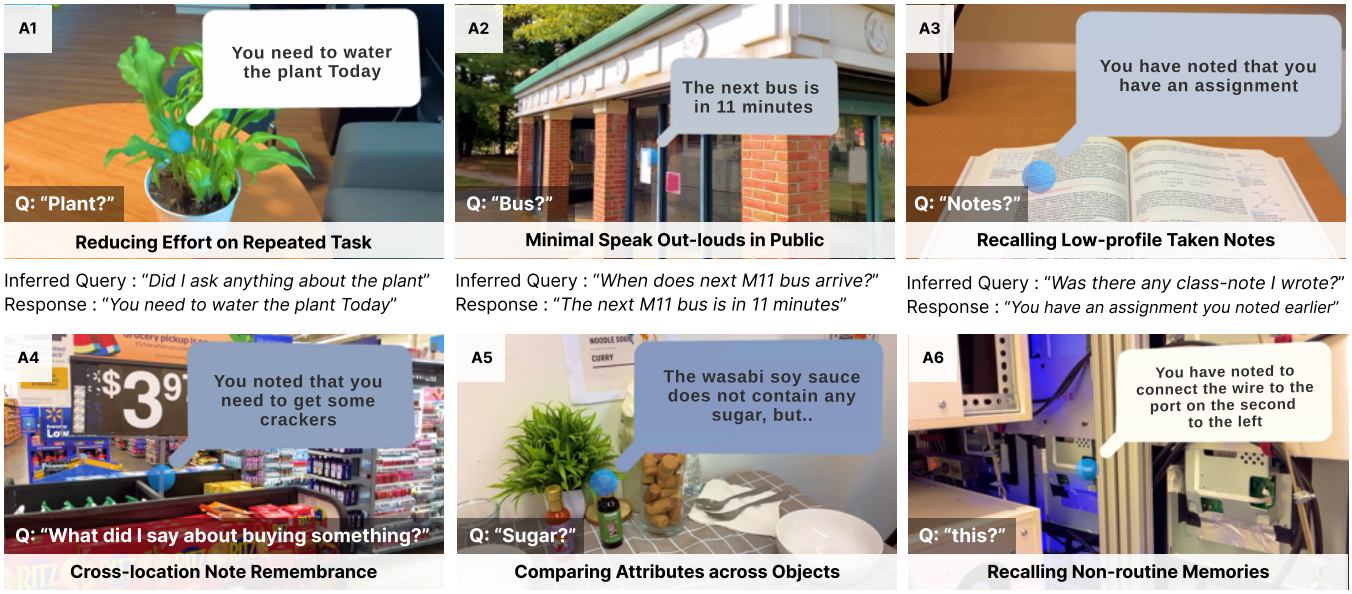

Figure 7: The six real-world use case scenarios of SpeechLess. The scenarios were inspired by its "in-the-wild" usages in Study B.

store, SpeechLess matches the intent of the query, retrieves the relevant memory, and surfaces the correct reminder.

5.5 Bistro: Within-space Cross-object Comparison

A user's nutritional content information comparison can be achieved without repeatedly articulating lengthy queries (A5). A user initially asks, *What is the sugar content of the wasabi soy sauce?*," gazing at the target. SpeechLess records the nutritional detail and the space as part of a memory. Later, while looking at a different sauce, *sugar?*" can recall the prior user interest at the Bistro, inferring the user's intent, and adaptively responds with the sugar content of the new sauce. This shows how SpeechLess connects the experience even across objects while lowering effort.

5.6 Maintenance: Non-routine Memory Recall

SpeechLess is effective in recalling non-daily memories too (A6). During maintenance, a verbal note of "*Remember to connect the wire to the port second from the left*." left a month ago, can simply be glanced and uttered "*this*" to recall the instruction notes. This is because the record of the spatial memory and its contextual cue is limited in number (uncommon visual features); retrieving the prior episode through SpeechLess, requires even less interaction cost.

6 Evaluation

We evaluate SpeechLess through two complementary studies. (1) Study A: a controlled lab study that assesses the benefits of spatially grounded memory and minimal-speech queries under counterbalanced conditions, and (2) Study B: an "in-the-wild" longitudinal deployment that examines the practical adoption of SpeechLess in everyday settings assessing its social acceptability and privacy concerns. Together, we provide preliminary evidence on the social dynamics of a speech-triggered intent resolution mechanism, and how it adapts under varying spatial context, with the research questions:

- **RQ1.** How does minimal/zero-utterance interaction influence human-AI interaction experience for daily routine activities?
- **RQ2.** What does spatial memory mean for a personal wearable assistant? Is there any special advantages of it?
- **RQ3.** Does the minimal use of verbal queries shift user behaviors when using the AI assistant in public?

We employed a within-subjects design with three utterance modes: *Full Utterance* (denoted "*Full*"), *Partial Utterance*, and *Zero Utterance* (denoted "*Zero*"). The *Full* acts as our baseline. Participants were informed of the data anonymization step, and provided consent prior to the participation. This study was conducted under the IRB approval of Stony Brook University (*1173920*).

6.1 Study A: Understanding the Articulation Burden

We conduct technical and user analyses of SpeechLess–accuracy, completion time, effort, perceived load, and usability under controlled lab setting. We concentrate on comparative analysis between vanilla AI assistance (*Full*) and micro (*Partial*, *Zero*), to address **RQ1** and **RQ2**. In this study, we solely evaluate the *interaction paradigm itself–speech-based intent granularity control and articulation effort*, thus, we use a controlled experimental AR platform (HMD), to *facilitate precise assessment and to isolate our mechanism from device ergonomics or form-factor effects*.

Study Design

Participants. We recruited 18 participants (12 male, 6 female), aged 21-25 ($\mu = 23.0$, $\sigma = 1.19$) with prior experience with XR ($\mu = 3.78$, $\sigma = 2.02$; 7-point Likert scale). Participants reported Full Professional Proficiency or higher in English: Full Professional (N=14), Native/Bilingual (N=4). They were compensated $15 for a 90-minute session. We denote them as P1 to P18.

Apparatus and Setup. We conduct our experiment on a Meta Quest 3 HMD in passthrough AR mode with our custom Google Cloud Platform-hosted server logging user interactions. We use a controlled experimental platform (HMD) to capture the users' response in a deterministic, programmable, and reproducible manner (access to camera, parameters, hardware buttons). This reduces variability from external factors such as gesture-detection errors, ensuring that observed effects solely reflect the performance of our proposed interaction mechanism. We map the Minimalistic Tap Interface to the Quest right controller 'A' button due to limited access to its haptic touch input on the HMD.

We leverage the same Memory Retriever of SpeechLess across all three conditions, as the current GPT API is limited in keeping the context within the conversation ("memory") by default, requiring the API invoker to manually manage it. To showcase the adaptability of SpeechLess under different space, time, intent, and objects,

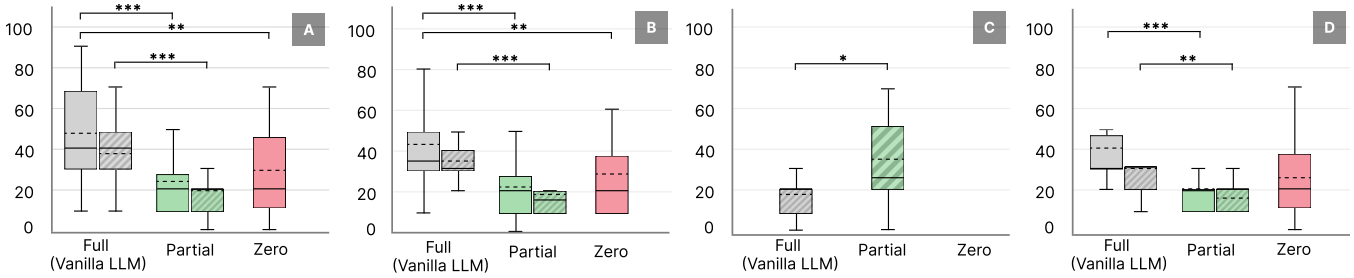

Figure 8: Comparison of participants' reported cognitive load with RTLX score : (A) Mental demand, (B) Physical demand, (C) Temporal demand, and (D) Effort; *** $p < .001$, ** $p < .01$, * $p < .05$; Full, Partial, Zero Utterance; Task1 (Non-striped), Task2 (Striped);

we made three personas, each indicating a hypothesized individual user, with disparate interests, and pre-seeded 45 personal memories before the study. This seed memory represents a hypothetical scenario where three distinct users have asked a number of questions prior to attempting to recall any memories. We construct three physical mock-up environments (Office, Break Room, and Bistro) to distribute these memories for a realistic use case scenario, and to study the adaptability of SpeechLess in spatial memory recalls. Refer to Supplementary Material for detailed study settings.

Tasks. Two controlled tasks were used to learn the interaction pattern of the participants with a conversational AI assistant across conditions, with referent grounding in realistic everyday contexts:

T1 - Remembrance: probes the recallability of previously seeded personal memories bound to objects and places (e.g., "*What did I say about* $\{topic\}$"). The task involves three physically different locations assessing spatial-memory aware information retrievals.

T2 - Comparison: probes information comparison between the object in the current view and another space or object (e.g., "*Can you compare the* $\{topic\}$ *between* $\{obj1\}$ *and* $\{obj2\}$"). The task evaluates an advanced use case of cross-space/object memory comparisons. *Zero* was not included in this task due to lacking cues to extrapolate user intent.

Procedure. Each participant received a brief introduction of the study procedure followed by a 20-minute tutorial, familiarizing them with the HMD and the functionalities of SpeechLess. The counterbalanced condition ordering and personas (45 seeded memories per persona) were given to the participant with the list of questions used during the memory seeding phase. The participants were given the freedom to randomly choose 18 (3 locations × 3+ trials × 2 task types) or more spatial memories across three different locations. Each participant was reminded to speak aloud and abide by the mode of interaction (*Full*, *Partial*, and *Zero*). After each Task trial, they completed a user experience survey: 0-100 scaled RTLX (Raw NASA-TLX), 7-point Likert scale survey, SUS (System Usability Scale [10]), and a semi-structured interview upon the completion of both Tasks. Repeated Measure ANOVA test (when normality held; Shapiro-Wilk test), or Friedman followed by Bonferroni corrected pairwise tests (or Wilcoxon signed-rank) was used.

User Evaluation

Perceived Relevance. Although no significance was identified across modes, participants reported that the answers were relevant to the spatial context. The relevance was measured with a 7-point scale (higher: more relevant). It showed that the mean score of *Full* was 5.94 (±0.87), followed by *Partial*, 5.44 (±1.04), and *Zero*, 5.11 (±1.68). Overall spatial adaptability (being able to recall memory relevant to a space) was rated 5.78 (±1.11), indicating significance above the neutral score of 4 (Wilcoxon, $p < .001$) and robust perceived recall across the three environments.

Cognitive Load. We report the user's perceived load across six RTLX sub-dimensions: Mental, Physical, Temporal, Performance, Effort, and Frustration (higher: more load; 0-100 scaled). Fig. 8 summarizes the distributions per condition and per significant pairs. Overall, abbreviating or removing articulation consistently reduced mental/physical/effort load without without harming perceived performance in T1, while in T2 the *Partial* condition lowered most loads but increased Temporal demand. No other dimensions showed statistical significance.

In T1, the *Mental demand* is ranked *Full* 47.78 (±24.87), *Zero* 29.44 (±27.96), and *Partial* 25.00 (±18.86); (Friedman. $\chi^2 = 23.41$, $p < .001$); and for pairwise-test, *Full* >*Partial* ($p < .001$), *Full* >*Zero* ($p < .01$). For both Physical demand and Effort, it follows the same trend. Physical demand: *Full* 42.78 (±22.18), *Zero* 28.33 (±23.33), and *Partial* 22.22 (±16.29) with *Full* >*Partial* ($p < .001$), *Full* >*Zero* ($p < .01$) ($\chi^2 = 19.00$, $p < .001$). Effort: *Full* 40.00 (±20.86), *Zero* 27.22 (±21.91), and *Partial* 21.11 (±11.83). However, in pair-wise analysis, only *Full* >*Partial* ($p < .001$) was significant ($\chi^2 = 20.22$, $p < .001$).

For T2, *Partial* reduced overall loads but increased temporal demand for some participants–without detailed keywords to allow SpeechLess to pin-point referred memories, it required additional attempts to retrieve the desired response: "*Full query mode didn't take much but Partial, I had to wait longer to process my query sometimes.*" (P10). The Mental demand score is ranked: *Full* 37.22 (±16.74) and *Partial* 19.44 (±12.59), ($p < .001$). And the Physical demand indicates: *Full* 34.44 (±10.42) followed by *Partial* 18.33 (±9.24), ($p < .001$). The Temporal demand: *Partial* 34.44 (±23.32) higher than *Full* 18.89 (±12.31), ($p = .025$). Finally, the Effort: *Full* 32.22 (±20.74) and *Partial* 16.67 (±6.86), $p = .006$.

Usability and Ease-of-use. While participants reported *Zero* as the most difficult approach to use (5.56±1.95) followed by *Full* (5.28±1.74) and *Partial* 4.78 (±1.59), the differences were not statistically significant across conditions. The mean perceived usability of the SpeechLess interface (across all conditions) was rated 'Good' with a SUS score of 75.42.

The usefulness for reducing articulation burden in *Partial* was scored 5.39 (±1.09) and *Zero*, 5.22 (±1.48), with significance compared to the neutral score of 4 (Wilcoxon, $p < .01$). Also, the use of *Partial* was evaluated easier to use compared to *Zero* due to its controllability in providing direct cues (verbal) and transparency: 5.67 (±1.08); (higher: ease-of-use preference of *Partial* over *Zero*) ($p < .001$); One of the participant said "*Zero was the most difficult, because I can't control it to reason in a certain way. It made the task feel harder and more mentally demanding*" (P17).

User Feedback and Preferences. In a questionnaire to rank the three conditions by preference (first being the most preferred), the result was evenly distributed among the conditions: *Partial* (N=7), *Full* (N=6), *Zero* (N=5). Our interview with the participants revealed positivity towards using *Partial* and *Zero* for convenience of use (P2, P5, P13), and *Full* for maximal explicitness (P1, P2, P4, P5, P13, P18). "*Full was more convenient for asking, Zero was easy as nothing to speak. Partial conveyed intents but sometimes annoying, I had to think of the most effective keyword*"(P5); and "*Partial was helpful in reducing effort while capturing my intents. Zero was great, I didn't have to say anything, it surprisingly knew.*" (P13)

Figure 9: Field test of SpeechLess on an established AR form factor.

Technical Evaluation

We collected 833 memories, each containing a dimension sketch, user query, and system response. We report latency, retrieval accuracy, and reduction in spoken effort.

Latency. From a question to a response, the mean (in seconds) latency of *Full* was 3.52 ($\pm$0.96)s followed by *Partial* 5.15 ($\pm$1.38)s, and *Zero* 2.42 ($\pm$1.04)s. *Zero* skips the transcription process as well as prior-based query intent analysis, directly inferring intent based on the visuo-spatial context, thus reaping a 31% increase in speed over *Full* (aligning with user feedback). On the other hand, *Partial* extrapolates missing intent specifics using prior memories on top of question answering latency, resulting in a high-delay response.

Accuracy. The mean rate of successful QA for *Full* was 95.4% ($\pm$2.1), while *Partial* inferred the intention behind the participants' query correctly, 86.7% ($\pm$3.4) of the time. *Zero* reaped 83.3% ($\pm$2.5) accuracy. *Zero* infers solely from the GPS location and visuals, not being able to capture intentions when there are more than one memory. The leading cause behind the *Full* errors was mispronounced questions, accent, or transcription failures.

Measured Effort. The average percentage decrease of word counts for *Partial*, compared to a *Full* query was 49.8% ($\pm$30.6). This indicates a reduced user effort in attempting to initiate a memory-based query compared to vanilla speech-based interaction. We omit the comparison to *Zero*, as it requires no spoken query.

6.2 Study B: Effects of SpeechLess in Everyday Public

This evaluation focuses on assessing the gains (or losses) that our *interaction paradigm induces in social contexts*. To examine system behavior under conditions where social discomfort is amplified, we conduct a longitudinal in-the-wild deployment. To evaluate the *isolated effects* of our speech-based interaction mechanism, we deploy SpeechLess on a well-established smart device form factor (iOS, Android). Learned insights from Formative Study indicate that smart-glasses form factors, as a newly introduced form factor, can *draw social attention in public settings, even in the absence of any explicit verbal interactions*. Such form factor-induced visibility introduces *novelty and social-attention effects that are orthogonal to the interaction mechanism of our evaluation*, limiting a fair assessment of the paradigm itself. By using mobile AR (socially familiar and widely accepted form factor), we *minimize these confounds* and *directly evaluate the system's ability to mitigate interaction friction* in everyday social environments (**RQ3**).

Study Design

Participants. We recruited 13 participants (9 male, 4 female), aged 22-35 ($\mu = 28.18$, $\sigma = 4.61$) with prior experience with XR ($\mu = 4.07$, $\sigma = 2.17$; 7-point Likert). Participants reported their level of extroverted-ness ($\mu = 4.01$, $\sigma = 1.34$). They were compensated $20 for a week-long participation. We denote them as P1 to P13.

System Setup. The functionalities of SpeechLess were maintained same as Study A, but ported to hand-held/mobile AR platforms in Unity. However, users were given two interfaces (software-level button) named "Ask" for *Full* and *Partial* queries, and "Silent Ask" for the *Zero* query (Fig. 9). Camera, Microphone, Network, and GPS sensors were accessed, and the app communicated with our Google Cloud Platform custom Python server.

Table 1: Qualitative coding of the "in-the-wild" diaries. The counts represent a comment (a comment can span across themes at once).

Themes	Learnable Insights
Social acceptability (N=18)	Default to *Partial/Zero* in socially sensitive contexts. Only fall back to *Full* when confidence is very low or disambiguation is inevitably required. Helps in public attention mitigation.
Public Privacy Concerns (N=7)	AR raising concerns due to a lack of established social-norms. People are unaware of what a new technology can do; I don't want to be stared at; The camera-like app in public drew more attention (misunderstood as taking a photo).
Appropriate for repeated tasks (N=14)	Realization of repeated use of smartphone during the study. Following similar pattern, daily, making use of the memory-based, repeating, with little effort. The app is simple. One interface for everything, no complicated conditions
Limited Use cases (N=10)	Used the app for routine tasks such as weather checking, and temperature; Yet, suffered from limited use cases. May require longer-term use of the app.
Queries with vague intent (N=19)	Ambiguous questions such as "based on my prior questions" or "a new generic question" leads to unexpected answers. Including a strong keyword such as "my" or "remember that.." helps clearly define intention.

Tasks and Guidelines. We maintained the task open-ended. Participants were instructed to "speak aloud" across diverse public and private settings, so that social comfort/awkwardness could be experienced. They were recommended to reach the daily quotas of 5 or more *Full* and 5 or more memory-based (*Partial* and *Zero*) queries. Participants were told to perform tasks alone (no peers) to experience social pressures. We investigate the natural use of our interaction paradigm on AR: the guidelines and tutorials were provided, but the usages were not strictly mandated, and use the app as often as they can in their routine. A Google Form daily diary was provided to be filled at the end of the day. The participants were invited to a private Slack channel for any study inquiries.

User Feedback and Experience

We thematically group the participants' feedback from their diary (N=57)–(Tab. 1), following the same procedure as Formative Study.

Social Acceptability and Benefits. Participants described *Partial/Zero* as useful in social silence-defined or dense-crowd venues. One participant described the improved user experience: "*It's uncomfortable having to talk in the store. I'd rather use Zero as much as possible until I need to speak to correct it.*" (P7). Pressure amplified (e.g., Grocery store, Airport, Library) the benefit of minimal/zero articulation. "*Everyone was quiet. I got some stares when I spoke aloud. So Zero was forced to be used*"(P5). In a 7-point Likert scale survey, micro/zero utterance substantially reduced the difficulty of "speaking aloud in public when alone." The mean public speaking difficulty (low:better) of *Full* being 4.28 ($\pm$1.82), micro interaction model was (*Partial* or *Zero*), 2.63 ($\pm$1.47).

Usability and Preferences. When asked to choose one mode for public/routine use, 63.2% of diary responses preferred *Partial/Zero* (N=36) over *Full* (N=21). The survey on the willingness to extend the capabilities of *Zero* or *Partial* function on a smart wearable device (e.g., glasses) was 5.10 ($\pm$1.81) (7-point scale), suggesting a positive tone with room for improvements. The primary reasoning behind the opposing view was due to the uncommonness of the form factor in society, "*people might stare at me*" (P2) and its lack of differentiating use cases from mobile platforms (Tab. 1). However, P2 indicated a strong need for a system such as SpeechLess, underscoring the importance of verbal privacy-protection and its design for public use. A participant (P5) evaluated highly its value for daily reminders and repetitive tasks with reduced effort (Tab. 1).

User Feedback. Overall, participants found micro-utterance useful. Especially for public scenarios, indicating positivity towards the angle of research. However, there were borderline cases where

Query Decoding Agent failed to parse the user's query intent between personal memory remembrance or a generic question inquiry. The natural language signals contain interpretation ambiguity (e.g., "*What is this?*"–can be a generic *Full* query, or may refer to user's personal meaning of the object), requiring nuanced, situational context to infer the query type. Another participant noted that the screen exposure of the mobile app to bystanders, increased public attention and reduced their comfortability when participating in public spaces such as a subway station. The participant requested a function that generates a 'proxy' of the moment of query [64], being able to concentrate on the content without having to worry about the AI assistant not being able to view the referent-of-question.

6.3 Takeaways and Analyses

Our studies showed that reducing articulation through *Partial* and *Zero* lowered perceived cognitive and physical effort without substantially degrading accuracy, validating the benefit of minimal and implicit intent expression (**RQ1**). Furthermore, "Spatial memory" served as a personal context repository and effectively bridged under-specified or absent speech with environmental cues, enabling robust intent extrapolation across locations and activities (**RQ2**). In everyday public settings, participants naturally adopted our low-articulation modes in socially sensitive contexts, and resorted back to the explicit mode (*Full*) when full intent clarification was needed. This demonstrates that a user-centric speech-based intent granularity control mechanism can effectively mitigate social friction while preserving expressive capacity (**RQ3**).

These findings derived across our Formative, Controlled, and Field assessments, highlight the value of a per-situation intent resolution strategy that adapts how intent is expressed based on contextual constraints and task complexity. Rather than privileging a single interaction mode, *we define the paradigm by treating speech as a flexible resource that users can regulate*, from *Full* articulation for explicit control, to *Partial* or *Zero* utterance when reduced exposure or effort is desired. Also, we emphasize the practical importance of wearers' own privacy management and agency, beyond bystander-oriented privacy mechanisms (e.g., video recording visual/audio indicators), through our evaluation.

7 Limitation and Discussion

Generality Across Wearable Form Factors. We validated our proposed interaction paradigm (speech-based intent granularity control with spatial memory) rather than the full ergonomics and social experience of a smart glasses form factor. Although our proposal is not medium-specific, smart glasses might reap different outcomes than an HMD or mobile (our study setup), such as different comfort thresholds for speaking, or stronger bystander attention. Thus, we plan to extend SpeechLess on smart glasses and replicate key measures under comparable routines to understand how form factor shifts the adoption boundary of the interaction paradigm.

Evolving Input Modalities Beyond Speech. We frame speech as the most expressive and currently dominant channel for conveying complex intent on contemporary wearables. However, emerging techniques [39, 62] and interfaces (e.g., Meta Neural Band [53]) may change the interaction landscape. Our design goal is not to privilege speech exclusively, but to provide a general mechanism for regulating how explicitly intent is externalized while completing the remaining intent dimensions from personal context. In our future work, we can leverage new modalities as additional low-cost intent signals that *complement* speech, while preserving the concept of the "granularity control + contextual completion" principle.

Recognizing Routine Patterns. The routine activity detection of SpeechLess currently relies on recency and word-based filters. However, the meaning of an action can shift substantially depending on its preceding context. More continuous modeling of user patterns is needed to capture everyday regularities. In future work, we will model repeated patterns beyond a single "dimension sketch" (space-time-referent-intent) by incorporating periodicity (daily/weekly) and semantic linking, enabling more nuanced retrieval and user needs prediction. We will also explore memory decay to mitigate long-term contextual overload [46].

Screen-space vs Situated Interface. We tackle the interaction paradigm for speech-controlled intent communication, a necessary step for wearable AR adoption, and allow users to have a manual choice between screen-space UI overlay and situated AR interface. However, practical wearable AR may require self-adaptive and context-awareness (e.g., topic-of-discussion, urgency, relevant referent, details, and occlusion [8,18,19,60,67,75]). We will extend our work towards an adaptive context-induced interface.

Arbitrating Multi-dimensional Contexts. SpeechLess resolves the fusion of spatiotemporal, intent, and verbal cues through Reciprocal Rank Fusion. While this method is functional, it lacks the robustness to handle nuanced conflicts, such as when a spoken reference contradicts a visual referent in view or when multiple dimensions suggest equally valid retrievals. Although designing an all-rounded recommender system is difficult, we plan to integrate dynamic inclusion/exclusion of dimensions by user's current context (parsing the needed "dimension sketch" using an agent), and add routine-pattern similarity cues.

Hardware Constraints and Sensing. We highlighted that current wearable devices impose practical constraints that shape what interaction paradigms are viable in everyday use–short battery life under continuous sensing, overheating, and limited on-device compute for sustained perception and LLM inference. These constraints also interact with privacy: continuous capture raises bystander concerns and increases the need for selective logging and filtering of non-essential information. We adopted an on-demand interaction model, prioritizing lightweight intent resolution over continuous lifelogging. As smart wearable domain shifts rapidly, including hardware and OS advances (e.g., Android XR [2]), we will revisit continuous, proactive information mining with the latest devices.

Towards Fully Proactive Wearable AR Assistance. While Contextual Dimension Encoder supports context-aware retrieval upon user initiation, we will extend SpeechLess toward proactivity by modeling predictive intent from user preferences (e.g., OmniActions [44], Sensible Agent [39]) and personalizing with learning-based methods (e.g., RL, SFT) using episodic memories and feedback from Verification Interface. We also plan to reduce micro-utterance latency by precomputing candidate intents, benefiting from emerging lightweight real-time LLMs.

8 Conclusion

SpeechLess introduced a proof-of-concept wearable AR assistant with a *speech-based intent granularity control paradigm* grounded in *personalized spatial memory*. By allowing users to regulate how explicitly intent is expressed (from full articulation to micro/zero-utterance), SpeechLess reduced articulation burden while reasonably preserving accuracy, usability, and user-agency, extending assistance to socially sensitive everyday settings. Through longitudinal, in-the-wild, and lab assessments, we demonstrated that our approach minimizes articulation and cognitive/physical effort while adapting everyday and socially constrained environments. We position SpeechLess as a foundational step toward scalable wearable interaction models that balance contextual intelligence, privacy, and expressive control, supporting long-term personalization and context-adaptive AR assistance [38,55].

Acknowledgments

We express our gratitude to Prantik Howlader, Nicholas Fraschilla-Brodkin, and Darius Coelho for their valuable feedback. This work was supported in part by NSF awards IIS2529207, IIS2441310 and ONR award N000142312124. Fig. 1 was edited using ChatGPT.

REFERENCES

[1] N. A. Akbar, R. Dembani, B. Lenzitti, and D. Tegolo. RAG-driven memory architectures in conversational llms-a literature review with insights into emerging agriculture data sharing. *IEEE Access*, 2025. 2

[2] Android. Android XR. https://www.android.com/xr/. Jan. 7. 2026. 9

[3] R. Arakawa, J. F. Lehman, and M. Goel. Prism-q&a: Step-aware voice assistant on a smartwatch enabled by multimodal procedure tracking and large language models. In *Proc. ACM IMWUT*, vol. 8, pp. 1–26, 2024. 2

[4] A. Assor, A. Prouzeau, M. Hachet, and P. Dragicevic. Handling non-visible referents in situated visualizations. *IEEE TVCG*, 30(1):1336–1346, 2023. 2

[5] S. A. Bahrainian and F. Crestani. Augmentation of human memory: Anticipating topics that continue in the next meeting. In *Proc. of ACM CHIIR*, pp. 150–159, 2018. 2

[6] L. Bajorunaite, S. Brewster, and J. R. Williamson. Virtual reality in transit: how acceptable is vr use on public transport? In *Proc. of IEEE VRW*, pp. 432–433, 2021. 2

[7] L. Bajorunaite, S. Brewster, and J. R. Williamson. Reality anchors: Bringing cues from reality to increase acceptance of immersive technologies in transit. *Proc. of ACM MHCI*, 7(MHCI), 2023. 2

[8] S. Boorboor, M. S. Castellana, Y. Kim, C. Zhu-tian, J. Beyer, H. Pfister, and A. E. Kaufman. VoxAR: adaptive visualization of volume rendered objects in optical see-through augmented reality. *IEEE TVCG*, 30(10):6801–6812, 2024. 9

[9] N. Bressa, J. Vermeulen, and W. Willett. Data every day: Designing and living with personal situated visualizations. In *Proc. of ACM CHI*, pp. 1–18, 2022. 2

[10] J. Brooke. Sus: A quick and dirty usability scale. *Usability Evaluation In Industry*, pp. 189–194, 1995. 7

[11] S. I. M. S. Bukhari, M. Sajid, B. Ji, and B. David-John. Rethinking privacy indicators in extended reality: Multimodal design for situationally impaired bystanders. In *Proc. of IEEE ISMAR-Adjunct*, 2025. 2

[12] W. Büschel, A. Lehmann, and R. Dachselt. Miria: A mixed reality toolkit for the in-situ visualization and analysis of spatio-temporal interaction data. In *Proc. of ACM CHI*, pp. 1–15, 2021. 2

[13] R. Cai, N. Janaka, H. Kim, Y. Chen, S. Zhao, Y. Huang, and D. Hsu. Aiget: Transforming everyday moments into hidden knowledge discovery with ai assistance on smart glasses. *arXiv:2501.16240*, 2025. 1

[14] R.-C. Chang, Y. Liu, and A. Guo. Worldscribe: Towards context-aware live visual descriptions. In *Proc. of ACM UIST*, pp. 1–18, 2024. 2

[15] Y. F. Cheng, A. Carden, H. Cho, C. G. Fidalgo, J. Wieland, and D. Lindlbauer. Augmented reality in-the-wild: Usage patterns and experiences of working with ar laptops in real-world settings. *arXiv preprint arXiv:2502.14241*, 2025. 2

[16] Q. Chu, H. Zhang, M. Liu, Y. Feng, H. Shi, and L. Nie. Intention-guided cognitive reasoning for egocentric long-term action anticipation. In *Proc. of AAAI*, 2026. 2

[17] M. Corbett, B. David-John, J. Shang, Y. C. Hu, and B. Ji. Bystandar: Protecting bystander visual data in augmented reality systems. In *Proc. of ACM MobiSys*, pp. 370–382, 2023. 2

[18] S. Davari and D. A. Bowman. Towards context-aware adaptation in extended reality: A design space for xr interfaces and an adaptive placement strategy. *arXiv preprint arXiv:2411.02607*, 2024. 1, 9

[19] S. Davari, F. Lu, and D. A. Bowman. Occlusion management techniques for everyday glanceable ar interfaces. In *Proc. of IEEE VRW*, pp. 324–330, 2020. 9

[20] M. D. Dogan, E. J. Gonzalez, K. Ahuja, R. Du, A. Colaço, J. Lee, M. Gonzalez-Franco, and D. Kim. Augmented object intelligence with XR-Objects. In *Proc. of ACM UIST*, pp. 1–15, 2024. 1, 2

[21] R. D. Easton and M. J. Sholl. Object-array structure, frames of reference, and retrieval of spatial knowledge. *JEP:LMC*, 21(2):483–500, 1995. 2

[22] J. Engel, K. Somasundaram, M. Goesele, A. Sun, A. Gamino, A. Turner, A. Talattof, A. Yuan, B. Souti, B. Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv:2308.13561*, 2023. 2

[23] C. M. Fang, Y. Samaradivakara, P. Maes, and S. Nanayakkara. Mirai: A wearable proactive ai" inner-voice" for contextual nudging. In *Proc. of ACM CHI EA*, 2025. 2

[24] P. Fung, Y. Bachrach, A. Celikyilmaz, K. Chaudhuri, D. Chen, W. Chung, E. Dupoux, H. Gong, H. Jégou, A. Lazaric, et al. Embodied ai agents: Modeling the world. *arXiv preprint arXiv:2506.22355*, 2025. 2

[25] Google. GoogleAI. Gemini models. https://ai.google.dev/gemini-api/docs/models/. Mar. 21. 2025. 5

[26] Google. Programmable search engine. https://developers.google.com/custom-search/v1/overview, 2025. Mar. 23. 2025. 5

[27] J. Grubert, T. Langlotz, S. Zollmann, and H. Regenbrecht. Towards pervasive augmented reality: Context-awareness in augmented reality. *IEEE TVCG*, 23(6):1706–1724, 2016. 1

[28] V. Y. Han, J. T. Gonzalez, C. Yang, Z. Wang, S. E. Hudson, and A. Ion. Towards unobtrusive physical ai: Augmenting everyday objects with intelligence and robotic movement for proactive assistance. In *Proc. of ACM UIST*, pp. 1–16, 2025. 2

[29] M. Harvey, M. Langheinrich, and G. Ward. Remembering through lifelogging: A survey of human memory augmentation. *PMCJ*, 27:14–26, 2016. 2

[30] Y. O. Hu, J. Tang, X. Gong, Z. Zhou, S. Zhang, D. S. Elvitigala, F. F. Mueller, W. Hu, and A. J. Quigley. Vision-based multimodal interfaces: A survey and taxonomy for enhanced context-aware system design. In *Proc. of ACM CHI*, pp. 1–31, 2025. 2

[31] S. Jang, E.-J. Ko, and W. Woo. Unified user-centric context: Who, where, when, what, how and why. In *Proc. of UbiPCMM*, 2005. 3

[32] M. S. U. Khan, M. Z. Afzal, and D. Stricker. SituationalLLM: proactive language models with scene awareness for dynamic, contextual task guidance. *arXiv:2406.13302*, 2024. 2

[33] O. Khan, Z. Ahmed, H. Nam, and K. Kim. TangibleMoments: Embedding XR memories onto physical objects. In *Proc. of IEEE VRW*, pp. 1147–1153, 2025. 2

[34] Y. Kim, Z. Aamir, M. Singh, S. Boorboor, K. Mueller, and A. E. Kaufman. Explainable XR: understanding user behaviors of XR environments using LLM-assisted analytics framework. *IEEE TVCG*, 31(5):1–11, 2025. 2, 3

[35] Y. Kim, S. Boorboor, A. Rahmati, and A. Kaufman. Design of privacy preservation system in augmented reality. In *Proc. of IEEE VizSec*, 2021. 2

[36] Y. Kim, S. Goutam, A. Rahmati, and A. Kaufman. Erebus: Access control for augmented reality systems. In *Proc. of USENIX Security*, pp. 929–946, 2023. 2

[37] R. K. Kundu, I. Ahmed, and K. A. Hoque. Pilar: Personalizing augmented reality interactions with llm-based human-centric and trustworthy explanations for daily use cases. *arXiv preprint arXiv:2512.17172*, 2025. 2

[38] B. Lee, M. Sedlmair, and D. Schmalstieg. Design patterns for situated visualization in augmented reality. *IEEE TVCG*, 30(1):1324–1335, 2023. 1, 9

[39] G. Lee, M. Xia, N. Numan, X. Qian, D. Li, Y. Chen, A. Kulshrestha, I. Chatterjee, Y. Zhang, D. Manocha, et al. Sensible agent: A framework for unobtrusive interaction with proactive ar agents. In *Proc. of ACM UIST*, pp. 1–22, 2025. 1, 2, 9

[40] J. Lee, J. Kim, J. Ahn, and W. Woo. Remote diagnosis of architectural heritage based on 5w1h model-based metadata in virtual reality. *ISPRS IJGI*, 8(8):339, 2019. 3

[41] J. Lee, J. Wang, E. Brown, L. Chu, S. S. Rodriguez, and J. E. Froehlich. GazePointAR: a context-aware multimodal voice assistant for pronoun disambiguation in wearable augmented reality. In *Proc. of ACM CHI*, pp. 1–20, 2024. 1, 2

[42] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS*, 33:9459–9474, 2020. 5

[43] C. Li, G. Wu, G. Y.-Y. Chan, D. G. Turakhia, S. Castelo Quispe, D. Li, L. Welch, C. Silva, and J. Qian. Satori: Towards proactive ar assistant

with belief-desire-intention user modeling. In *Proc. of ACM CHI*, pp. 1–24, 2025. 1, 2

[44] J. N. Li, Y. Xu, T. Grossman, S. Santosa, and M. Li. OmniActions: predicting digital actions in response to real-world multimodal sensory inputs with LLMs. In *Proc. of ACM CHI*, pp. 1–22, 2024. 2, 9

[45] J. N. Li, Z. J. Zhang, and J. Ma. Omniquery: Contextually augmenting captured multimodal memory to enable personal question answering. In *Proc. of ACM CHI*, 2025. 2

[46] T. Li, L. Jin, Z. Wu, and Y. Chen. Combined recommendation algorithm based on improved similarity and forgetting curve. *Information*, 10(4):130, 2019. 9

[47] J. Liu, K. A. Satriadi, B. Ens, and T. Dwyer. Investigating the effects of physical landmarks on spatial memory for information visualisation in augmented reality. In *Proc. of IEEE ISMAR*, pp. 289–298, 2024. 2

[48] X. B. Liu, S. Fang, W. Shi, C.-S. Wu, T. Igarashi, and X. Chen. Proactive conversational agents with inner thoughts. In *Proc. of ACM CHI*, 2025. 1

[49] L. Long, Y. He, W. Ye, Y. Pan, Y. Lin, H. Li, J. Zhao, and W. Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. *arXiv preprint arXiv:2508.09736*, 2025. 2

[50] F. Lu and D. A. Bowman. Evaluating the potential of glanceable ar interfaces for authentic everyday uses. In *IEEE VR*, pp. 768–777, 2021. 2

[51] F. Lu, L. Pavanatto, and D. A. Bowman. In-the-wild experiences with an interactive glanceable ar system for everyday use. In *Proc. of ACM SUI*, pp. 1–9, 2023. 2

[52] Z. Lv, N. Charron, P. Moulon, A. Gamino, C. Peng, C. Sweeney, E. Miller, H. Tang, J. Meissner, J. Dong, et al. Aria everyday activities dataset. *arXiv:2402.13349*, 2024. 2

[53] Meta. EMG Wristbands and Technology. https://www.meta.com/emerging-tech/emg-wearable-technology/. Jan. 7. 2026. 9

[54] C. Meurisch, C. A. Mihale-Wilson, A. Hawlitschek, F. Giger, F. Müller, O. Hinz, and M. Mühlhäuser. Exploring user expectations of proactive ai systems. *Proc. of ACM IMWUT*, 4(4):1–22, 2020. 1

[55] P. Milgram and F. Kishino. A taxonomy of mixed reality visual displays. *IEICE TIS*, 77(12):1321–1329, 1994. 9

[56] L. Ning, L. Liu, J. Wu, N. Wu, D. Berlowitz, S. Prakash, B. Green, S. O'Banion, and J. Xie. User-llm: Efficient llm contextualization with user embeddings. In *Proc. of ACM WWW*, pp. 1219–1223, 2025. 2

[57] A. Paruchuri, S. Hersek, L. Aggarwal, Q. Yang, X. Liu, A. Kulshrestha, A. Colaco, H. Fuchs, and I. Chatterjee. Egotrigger: Toward audio-driven image capture for human memory enhancement in all-day energy-efficient smart glasses. *IEEE TVCG*, 2025. 1

[58] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos. Context aware computing for the internet of things: A survey. *IEEE Commun. Surv. Tutor.*, 16(1):414–454, 2013. 2

[59] K. Pu, T. Zhang, N. Sendhilnathan, S. Freitag, R. Sodhi, and T. R. Jonker. Promemassist: Exploring timely proactive assistance through working memory modeling in multi-modal wearable devices. In *Proc. of UIST*, pp. 1–19, 2025. 1, 2

[60] A. Raianova and M. Lee. Adaptive learning in extended reality: A survey on multimodal interaction and ai-driven personalization. In *Proc. of IEEE ISMAR-Adjunct*, pp. 205–210, 2025. 9

[61] S. Rajaram, M. Peralta, J. G. Johnson, and M. Nebeling. Exploring the design space of privacy-driven adaptation techniques for future augmented reality interfaces. In *Proc. of ACM CHI*, pp. 1–19, 2025. 2

[62] S. Rajaram, H. B. Surale, C. McConkey, C. Rognon, H. Mehta, M. Glueck, and C. Collins. Gesture and audio-haptic guidance techniques to direct conversations with intelligent voice interfaces. In *Proc. of ACM CHI*, pp. 1–20, 2025. 1, 2, 3, 9

[63] L. Rau, J. L. Bitter, Y. Liu, U. Spierling, and R. Dörner. Supporting the creation of non-linear everyday ar experiences in exhibitions and museums: An authoring process based on self-contained building blocks. *Front. Virtual Reality*, 3:955437, 2022. 2

[64] K. A. Satriadi, A. Cunningham, R. T. Smith, T. Dwyer, A. Drogemuller, and B. H. Thomas. Proxsituated visualization: An extended model of situated visualization using proxies for physical referents. In *Proc. of ACM CHI*, pp. 1–20, 2023. 9

[65] K. A. Satriadi, B. Tag, and T. Dwyer. Context-dependent memory in situated visualization. *arXiv:2311.12288*, 2023. 2

[66] J. Shen, J. J. Dudley, and P. O. Kristensson. Encode-store-retrieve: Augmenting human memory through language-encoded egocentric perception. In *Proc. of IEEE ISMAR*, pp. 923–931, 2024. 1, 2

[67] E. Song, T. Ha, J. Park, H. Lee, and W. Woo. Holistic quantified-self for context-aware wearable augmented reality. *IJHCS*, p. 103568, 2025. 9

[68] D. Stover and D. Bowman. Taggar: General-purpose task guidance from natural language in augmented reality using vision-language models. In *Proc. of ACM SUI*, pp. 1–12, 2024. 2

[69] T. T. M. Tran, S. Brown, O. Weidlich, S. Yoo, and C. Parker. Wearable ar in everyday contexts: Insights from a digital ethnography of youtube videos. In *Proc. of ACM CHI*, 2025. 2

[70] Y. Wang, Y. Lu, S. Yan, and X. Shen. "If My Apple Can Talk": Exploring the use of everyday objects as personalized ai agents in mixed reality. In *Proc. of ACM CHI EA*, pp. 1–9, 2025. 2

[71] X. Xu, A. Yu, T. R. Jonker, K. Todi, F. Lu, X. Qian, J. M. Evangelista Belo, T. Wang, M. Li, A. Mun, et al. Xair: A framework of explainable ai in augmented reality. In *Proc. of ACM CHI*, pp. 1–30, 2023. 1, 3, 5

[72] B. Yang, L. Xu, L. Zeng, K. Liu, S. Jiang, W. Lu, H. Chen, X. Jiang, G. Xing, and Z. Yan. ContextAgent: Context-aware proactive llm agents with open-world sensory perceptions. *NeurIPS*, 2025. 1

[73] J. Yang, S. Yang, A. W. Gupta, R. Han, L. Fei-Fei, and S. Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proc. of IEEE/CVF CVPR*, pp. 10632–10643, 2025. 2

[74] X. Zhang, Y. Deng, Z. Ren, S. K. Ng, and T.-S. Chua. Ask-before-plan: Proactive language agents for real-world planning. In *Prof. of ACL EMNLP*, pp. 10836–10863, 2024. 2

[75] J. Zheng, H. Weng, X. Wang, C. Cui, S. Mayer, C.-L. Tai, and L.-H. Lee. Persono: Personalised notification urgency classifier in mixed reality. In *Proc. of IEEE ISMAR*, pp. 1053–1063, 2025. 9

[76] C. Zhu, S.-K. Hsia, X. Hu, Z. Liu, J. Shi, and K. Ramani. agentar: Creating augmented reality applications with tool-augmented llm-based autonomous agents. In *Proc. of ACM UIST*, pp. 1–23, 2025. 2

[77] W. D. Zulfikar, S. Chan, and P. Maes. Memoro: Using large language models to realize a concise interface for real-time memory augmentation. In *Proc. of ACM CHI*, pp. 1–18, 2024. 2