

# Understanding Spatiotemporal-Aware Multimodal Conversational Search for Use in the Outdoor Urban Space

Jiangnan Xu  
 jiangnan.xu@tuni.fi  
 Rochester Institute of Technology  
 Rochester, NY, USA  
 Tampere University  
 Tampere, Finland

Suyeon Seo  
 suyeonseo@yonsei.ac.kr  
 Yonsei University  
 Seoul, Korea

Joni Salminen  
 jonisalm@uwasa.fi  
 University of Vaasa  
 Vaasa, Finland

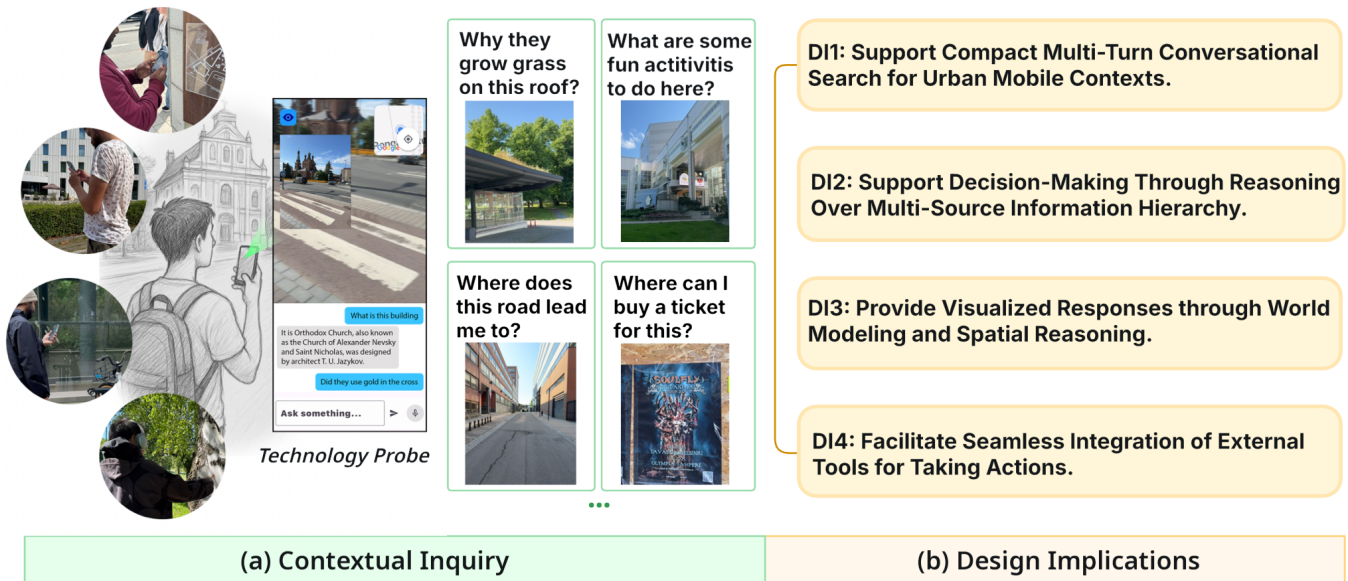
Michael Saker  
 Michael.Saker@city.ac.uk  
 City, University of London  
 London, UK

Joongi Shin  
 joongi.shin@aalto.fi  
 Aalto University  
 Espoo, Finland

Alan Chamberlain  
 Alan.Chamberlain@Nottingham.ac.uk  
 University of Nottingham  
 Nottingham, UK

Konstantinos Papangelis  
 kxpigm@rit.edu  
 Rochester Institute of Technology  
 Rochester, NY, USA

Dae Hyun Kim\*  
 dhkim16@yonsei.ac.kr  
 Yonsei University  
 Seoul, Korea  
 POSTECH  
 Pohang, Korea



**Figure 1: The overview of our study: (a) a contextual inquiry study with 23 urban dwellers using a technology probe (UrbanSearch) in a mobile urban context, and (b) four resulting design implications (DIs) for designing future multimodal conversational search (MCS) systems in the outdoor urban space.**

\*Corresponding author

CHI '26, April 13–17, 2026, Barcelona, Spain  
 © 2026 Copyright held by the owner/author(s).  
 This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*, April 13–17, 2026, Barcelona, Spain, <https://doi.org/10.1145/3772318.3790541>.

## Abstract

Emerging multimodal conversational search (MCS) tools (e.g., Gemini Live) allow users to search for spatiotemporal information through natural language dialogues as they move through urban space. Despite the growing popularity of these tools, there is limited understanding of how people engage with this technology. To address this gap, we developed *UrbanSearch*, an MCS technology

probe designed to capture the user's current geolocation, time, and visual surroundings. A contextual inquiry (N=23) revealed that MCS tools provide two core values: requiring low effort in forming queries while offering highly relevant responses, and functioning as a central information gateway. As a promising technology, MCS supports environmental learning, in-situ decision making, and personalized navigation. Participants also revealed unmet needs for spatial reasoning and transparent integration of multi-source information, along with concerns related to peripheral awareness, social context, and personal space. Drawing from the findings, we discuss design implications for future MCS tools in urban spaces.

## CCS Concepts

Human-centered computing; Empirical studies in HCI; Natural language interfaces.

## Keywords

Urban Space, Conversational Search, Contextual Inquiry

### ACM Reference Format:

Jiangnan Xu, Suyeon Seo, Joni Salminen, Michael Saker, Joongi Shin, Alandoor Chamberlain, Konstantinos Papangelis, and Dae Hyun Kim. 2026. Understanding Spatiotemporal-Aware Multimodal Conversational Search for Use in the Outdoor Urban Space. In Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13 17, 2026, Barcelona, Spain. ACM, New York, NY, USA, 20 pages. <https://doi.org/10.1145/3772318.3790541>

## 1 Introduction

People actively seek everyday information on mobile devices as they move through outdoor urban spaces [5, 49, 79, 90]. Common examples include choosing a place to eat [26] or learning the history of a nearby building [5, 45]. Yet in transit [78], people often struggle to locate relevant information sources through traditional web search [13, 49] due to scarce attention [20, 78] and single-turn queries that under-specify evolving, situated needs [66]. To address these needs, emerging multimodal conversational search (MCS) tools (e.g., SearchGP [67], Gemini Live [28]), powered by large language models (LLMs), present promising alternatives [57]. Technologies such as artificial intelligence (AI) and mobile computing provide new ways of integrating vision models with geographical data. For example, through its integration with Google Maps [3], Gemini Live allows users to ask questions about their surroundings directly via their smartphone cameras, while augmented reality (AR) glasses (e.g., [34]) promise to overlay real-time, location-aware information directly into users' fields of view as they navigate.

MCS tools facilitate search interaction by supporting multi-turn natural language dialogues and accepting both voice/text and visual input, which helps users clarify their intent more effectively [58]. However, most research on conversational search [58, 66] has focused on stationary and indoor contexts (e.g., home or office), leaving human interactions with MCS tools in outdoor urban spaces underexplored. Although these emerging applications demonstrate growing capabilities for spatiotemporal-aware [35] MCS interaction, the human-computer interaction (HCI) community lacks empirical studies of people's behaviors, unmet needs, concerns, and expectations in MCS for use in outdoor urban spaces. This matters

because the information search experience in outdoor urban spaces differs crucially from indoor stationary contexts, as users must respond to dynamic surroundings. There are specific behaviors such as location-based and contextualized queries [22, 37] that necessitate search tools capable of interpreting spatiotemporal context. For example, to answer the query, "What are some fun activities to do here?", the system must incorporate both spatial context (e.g., geographical location, visual surroundings) and temporal context (e.g., time of day, season). Understanding these conditions is vital for the design of human-centered MCS systems.

Despite the potential convenience introduced by MCS technology, there is growing attention to its risks [1], such as information trust [44, 77], and data privacy [7, 36]. Even though recent work acknowledges these issues, tensions and concerns specific to outdoor urban spaces remain underexplored. For example, public settings involve bystanders and social norms, suggesting that MCS technology affects not only individual users but also other urban residents [4].

These dynamics warrant further investigation, motivating the current work which aims to better understand how users envision both the potential and concerns of the MCS technology in the outdoor urban space. In particular, we put forth the following research questions (RQs):

RQ1: How would people interact with a spatiotemporal-aware MCS tool in the outdoor urban space?

RQ2: What expectations and concerns do people have about using such technology in the outdoor urban space?

To address these RQs, we designed and developed UrbanSearch, an LLM-powered MCS tool that serves as a technology probe in our study. UrbanSearch features a vision language model and web search capabilities, accepts user queries via text or audio, incorporates captured camera-view images, and delivers responses in both text and audio. To support spatiotemporal awareness, UrbanSearch also records the user's current GPS location and timestamp, which are used to interpret queries and generate contextually relevant responses. To use UrbanSearch for addressing our RQs, we conducted a real-world contextual inquiry in an urban environment, observing 23 people living in the city (from 2 weeks to 36 years) as they used UrbanSearch during walking sessions, and complemented these observations with in-depth semi-structured interviews to further investigate participants' experiences.

Our findings show that MCS in outdoor urban spaces is shaped by shifting attention and dynamic surroundings, unfolding within topic-varied and fragmented user behaviors, which we conceptualize as Observe Decide Act flows. Our findings indicate that MCS tools provide two core values: requiring low effort in forming queries while offering highly relevant responses, and functioning as a central information gateway. Participants envisioned spatiotemporal-aware MCS tools as enhancing urban life by supporting environmental learning, in-situ decision making, and personalized navigation, offering capabilities that conventional search engines do not provide. At the same time, they reported unmet needs, particularly around spatial reasoning (see Figure 5). In addition, participants raised concerns about using MCS tools in public settings, especially regarding safety, social awareness, and privacy.

In sum, our work makes the following contributions:

We provide empirical insights into how urban dwellers interact with a multimodal conversational search tool in an urban mobile

context. The findings reveal, on the one hand, the values MCS offers for on-the-move information seeking and, on the other hand, tensions that differ from stationary search contexts, including conflicts between observed surroundings and retrieved information, as well as unmet needs for spatial reasoning. We derive four design implications for spatiotemporal-aware MCS tools (Figure 1). These implications emphasize the need for compact multi-turn conversational search, spatiotemporal relevance reasoning across information sources, spatial reasoning capability, and information interoperability with external applications for action taking.

## 2 Related Work

In this section, we present related work of our research: (1) information seeking in urban mobile context, (2) multimodal conversational search, and (3) using intelligent systems in urban spaces.

### 2.1 Information Seeking in Urban Mobile Context

Urban space is a dynamic construct shaped by societal influences, encompassing physical environments that reflect political, cultural, and economic factors. It evolves continuously, adapting to changing social circumstances and human interactions [30]. Scholars have described urban spaces as information fields, where visual, tactile, and auditory cues guide pedestrian movements [68] and as relational assemblages in which communication and spatial experience are deeply intertwined [56, 65]. Beyond supporting wayfinding, technologies in urban space also mediate how people interpret, negotiate, and reimagine the city itself.

For example, Warner et al.'s work [5], echoing situationist perspectives [27, 40], demonstrates that walking in the urban environment can be designed not only for efficiency but also to surface alternative values, encourage reflection, and shape the experiential qualities of movement. Through walking, urban residents engaged with the city in ways that extended beyond goal-oriented navigation, encountering new social, political, and historical ambiances that reframed how urban space was experienced [5]. At the same time, information search in urban spaces has shifted from fixed infrastructures to mobile practices [37]. Fixed information infrastructures in cities, such as public displays [62, 95] and sensor-based installations [2, 38], have traditionally mediated how people access and share information in public environments. With the prevalence of smartphones, however, mobile information seeking has become increasingly common [72]. People now access information wherever they are [17], which is influenced by diverse physical and social contexts, interwoven with movement, embodied experience, and environmental stimuli [72].

Importantly, mobile information search [1] is a broader concept encompassing information searching on the move. Mobile information search refers broadly to accessing information via mobile devices and wireless internet, encompassing both use contexts while stationary and while on the move [8, 24, 46], while information search on the move focuses specifically on highlighting the user as an active moving entity and emphasizes how the dynamic physical environment shapes the information searching process [36, 78]. Our study focuses on the latter context: information search on the move.

Large-scale studies of mobile search behavior have shown that smartphones shape search practices closely tied to users' immediate environments [24, 74], which distinguishes information search on the move from other information search behaviors in everyday life [69], especially from search in stationary indoor contexts (e.g., library [4], home [53]). Yet prior research often centers on mobile device usage [1, 72], paying less attention to the dynamic physical environment as an active element. As information searching on the move becomes increasingly prevalent [37, 72], this technology use case merits dedicated investigation. In particular, the specific dynamics of moving through urban space where information needs are triggered by visual stimuli, spatial transitions, and situational awareness [1] remain underexplored. We argue that mobile information seeking is shaped not only by the smartphone as an embodied device [7] but also by the dynamic environment in which it is used, making the urban space itself an active element in the information search experience. To this end, our work investigates mobile information seeking in outdoor public urban spaces, where people are conditioned by mobility, environmental triggers, and fragmented attention [39].

### 2.2 Multimodal Conversational Search

Conversational search [58, 97] enables users to seek information through multi-turn dialogue, offering a promising alternative to traditional keyword-based web search for more complex and precise information retrieval. Conversational search allows users to refine ambiguous queries, clarify intent, and obtain contextually relevant results [100]. With the advent of LLMs, conversational search has become significantly more capable and accessible, as demonstrated by tools such as ChatGPT [58]. At the same time, vision language models (VLMs) [5] extend conversational systems to handle multimodal inputs such as images and live video, signaling a growing trend toward MCS.

Beyond visual awareness, integrating spatiotemporal awareness [5] shows further potential for supporting information seeking that is grounded in where and when a query arises. We define spatiotemporal awareness as a system's ability to situate queries in both temporal and spatial contexts, and in our work, we conceptualize this as an MCS tool that integrates and processes spatial context (e.g., location, visual surroundings) and temporal context (e.g., time of day, season). Recent technical attempts, such as Gemini [28] and AR glasses [34, 73], demonstrate early steps toward this vision by allowing users to ask questions about their environment while walking in the city. However, current applications show little evidence of robustly integrating and reasoning with location, time, and visual context, and in parallel, research has yet to establish what users actually need or expect from such capabilities. This two-sided gap leaves open questions about how people would engage with spatiotemporal-aware MCS in the real world and what kinds of features these systems should provide. In short, the HCI community lacks knowledge of spatiotemporal-aware MCS: what its full potential could be, or where its practical limits lie, which leaves open critical questions about how users would engage with such systems and what kinds of features they would truly value.

### 2.3 Using Intelligent Systems in Urban Spaces

Prior work has demonstrated the potential of AI systems to mediate human experiences in urban spaces, from offering practical assistance to enabling playful and reflective interactions [42, 45]. For example, Runze et al. developed AiGet, a system that leverages wearable eye-tracking and AI to support informal learning during everyday walks [8]. Hung et al. [42] explored how generative AI can be used playfully in urban environments. While these studies demonstrate the potential of AI as an active mediator between people and urban spaces, offering new ways to explore, understand, and engage with surroundings, HCI scholars have identified risks associated with the deployment of AI in everyday contexts. Weidinger et al. [91] outline the broad ethical and social risks posed by LLMs, including misinformation, bias, and potential harms from automation at scale. Gümürel et al. [36] focus specifically on user privacy, proposing a framework to analyze the types of harms conversational AI may cause in everyday use. Similarly, Ali et al. [7] examine users' own attitudes, concerns, and expectations toward security and privacy in conversational AI platforms. While these works pinpoint important ethical, privacy, and social concerns, it remains underexplored how such concerns manifest when people engage with AI-assisted systems in urban spaces, where interactions are intertwined with mobility, public settings, and situational social context [1]. In turn, these considerations motivate our work to empirically probe how people might use MCS in urban spaces, surfacing real-world practices, unmet needs, and concerns to inform future design.

### 3 Method

To explore how users engage in situated, conversational search while moving through the outdoor urban space, we developed a mobile application, UrbanSearch, as a technology probe. Using contextual inquiry, a field method that combines observation and interviewing in natural settings to uncover real-world behaviors [12], we observed participants' use of UrbanSearch during an outdoor walking session in urban spaces, and conducted in-depth, semi-structured interviews to reflect on participants' experiences. We conducted this study in Tampere, the second-largest city in Finland. To deepen our understanding of user behavior and perception, we employed a triangulated data analysis approach [11], drawing from multiple data sources including probe logs, screen recordings, and interview transcripts.

#### 3.1 Technology Probe: UrbanSearch

Guided by Hutchinson et al.'s technology probe framework [43], we designed and developed UrbanSearch (see Figure 2). UrbanSearch combines visual-text query support with spatiotemporal awareness to investigate how locative and real-time contexts affect user interactions. We began with geolocation, time, and visual surroundings as the basis for spatiotemporal awareness, since they shape common information seeking in urban spaces such as "what is nearby?", "what is open now?", or "what is this?" [28, 72]. Developing our own technology probe allowed detailed logging of user input and probe output to support analysis and identify unmet needs, informing future design directions [43].

**3.1.1 User Interface.** UrbanSearch enables conversational search across multiple modalities. Users can type queries into a text field or

Figure 2: UrbanSearch, the technology probe interface. The interface supports multimodal interaction for situated information seeking in urban space. Users can capture their camera view (top-left) and attach it with a text query, and check their current location in Google Maps through the map shortcut (top-right). Conversational search occurs through the conversational interface (bottom), which accepts text or audio input and returns text + audio responses in a multi-turn dialogue format.

or speak by holding a microphone button, with speech transcribed to text. The camera provides a live view of the surroundings without recording by default, and users can tap the eye icon to capture a photo to accompany a text query. Users can also see a mini-map that displays their current location. Tapping the mini-map launches Google Maps, enabling them to navigate if needed. Although UrbanSearch does not provide built-in navigation, this integration supports fluid transitions between query and movement. Users interact through a chat-like interface, where responses are delivered both as on-screen text and spoken aloud using text-to-speech. The full conversation appears as a scrollable message list, allowing users to follow along or revisit past conversations.

**3.1.2 Query Response Generation.** Upon receiving a query, UrbanSearch gathers real-time contextual information. This includes the user's current GPS coordinates, the current timestamp, and a camera image if the user chooses to capture one. To obtain the user's address and nearby places, we combined their GPS coordinates with the Google Maps API for reverse geocoding. To enrich the response with up-to-date data, the probe also performs web searches using an LLM tailored for handling web search queries (gpt-4o-search-preview) via an OpenAI API. UrbanSearch then composes a structured prompt (see Supplementary Material) that

<sup>1</sup>We followed iOS location privacy requirements, asking for user permission before accessing location data. We informed the participants and asked for their consent regarding the collection of location data prior to the study.

integrates spatial, temporal, visual, and web-searched results along with the conversation history. Importantly, we tailored the system prompt to foreground spatiotemporal awareness when generating the response. The probe submits this prompt to an LLM (40) via the OpenAI API and returns a generated response tailored to the user's input query. The backend server (hosted on Railway) logs all user input turns (text queries, captured images) and UrbanSearch responses. These logs serve two purposes: (1) to support reflection and discussion during user study interview sessions (see Section 3.3.3) by using each user's data, and (2) to enable data analysis across all participants (see Section 3.4).

## 3.2 Participants

After receiving an Institutional Review Board (IRB) approval, we recruited participants in Tampere, including both short-term visitors and long-term residents, via local mailing lists and social media. The recruitment post briefly described the study purpose and listed inclusion criteria: aged 18–65, fluent in English, available for a two-hour in-person user study, and comfortable walking in outdoor urban environments for around 30 minutes while using a mobile application. Individuals interested in participation completed a pre-screening survey, which collected demographic information (age, gender, professional background), self-reported familiarity with AI search tools (i.e., conversational agents and AI search engines, measured on a 5-point Likert scale), weekly frequency of AI search tool use, commonly used AI search tools (e.g., ChatGPT, Gemini, Copilot, Google Lens), as well as their familiarity with four city center areas respectively (measured on a 5-point Likert scale) and the duration of their presence in the city. To choose an appropriate city area for the study where participants would have mixed levels of familiarity, we selected one of the four areas that showed the most balanced range of familiarity. To ensure a diverse participant pool, we employed a rolling recruitment strategy with purposeful sampling, prioritizing underrepresented profiles mainly with respect to participant familiarity with the study area and AI search tools, but also with respect to their demographics.

In total, we recruited 23 participants (11 female, 9 male, 2 non-binary, 1 prefer not to say), aged between 20 and 57 years ( $M = 31.8$ ,  $SD = 10.5$ ). As shown in Table 1, participants had a range of familiarity with AI search tools (Column 4) and exposure to AI search tools (0 to more than 5 times per week; Column 5), as well as familiarity with the specific study area (Column 6) and duration of presence in the city (2 weeks to 36 years; Column 7).

## 3.3 Contextual Inquiry Procedure

We conducted the contextual inquiry study one-on-one, with the first author accompanying each participant. For each participant, we provided an iPhone 16 Pro device with UrbanSearch installed. Each session consisted of three parts: (1) an introduction and tutorial (about 10 minutes), (2) an in-the-wild technology probing session (about 30 minutes), and (3) a post-session semi-structured interview (about 50 minutes). We include the complete study protocol and interview questions in the Supplementary Material. Each study session lasted approximately 1.5 hours in total. Upon completion of the study, each participant received a \$30 USD e-gift card as compensation for their time.

**3.3.1 Introduction and Tutorial Session.** In the introduction and tutorial session, we explained the background and the overall procedure of the study, and received data collection consent from the participants. We explained that the aim of the study was to explore how users interact with a spatiotemporal-aware MCS tool in outdoor urban spaces and reflect on unmet needs and future possibilities, rather than to evaluate UrbanSearch as a product [40]. We then walked the participant through the user interface and demonstrated how to use UrbanSearch. To serve as a context for the contextual inquiry (e.g., find a fruit market, visit a famous landmark), we asked participants to start the walking session in a way that felt natural to them: either exploratory strolling or a specific initial goal in mind. We provided some sample goals based on prior work (finding a restaurant for a meal [79], learning about a nearby non-human object (e.g., architecture, sculpture [45]) for those desiring to start with an initial goal. To maintain ecological validity, we told the participants that they were free to use other applications as needed (e.g., web search) and could access Google Maps through the probe's interface. We also informed the participant that UrbanSearch might hallucinate or provide unsatisfactory responses.

**3.3.2 Technology Probing Session.** Participants used the probe while walking outdoors for approximately 30 minutes. We showed the rough area they can freely move through in a map (see Figure 8 in Appendix B). We selected this duration to allow sufficient engagement with UrbanSearch while avoiding user fatigue. Two prior pilot studies indicated that thirty minutes was an appropriate amount of time for participants to complete initial onboarding and show natural use of UrbanSearch, given most participants had prior experience of using AI search tools such as ChatGPT. We informed participants they could briefly pause or engage in small activities beyond using the probe in outdoor spaces (e.g., buying food from a truck), but clarified that longer activities, such as indoor dining or extended shopping, were outside the scope of the session. The first author followed each participant at a distance to ensure safety and took observational notes. We asked participants to think aloud, reflecting on their behaviors and experiences in real time.

**3.3.3 Post-session Interview.** After the outdoor session, we went indoors for an in-depth, semi-structured interview. To ground the discussion in the interview, we presented the participant with their screen recording, captured images, and a spreadsheet table of their query logs. This enabled the participant to revisit their interactions, reflect on specific inputs and outputs, and elaborate on their intentions, expectations, and reactions during the session. The interview consisted of three segments: (1) Overall experience reflection, in which we asked participants to describe their overall experience, moments of confusion, and interactions they found meaningful or surprising; (2) Query-level review, in which we guided participants through their input-output query logs and screen recordings, prompting them to comment on interactions they liked, disliked, or found unexpected, and to articulate why and what they had hoped UrbanSearch would have done; and (3) Expectations and future use, in which we asked participants for their expectations for using the MCS technology, possible usage scenarios in everyday life, concerns about the MCS technology, and suggestions for improvement.

Table 1: Summary of contextual inquiry participants. Gender categories: F = Female, M = Male, NB = Non-binary, X = Prefer not to say. Familiarity is measured on a 5-point scale (5 = very familiar, 1 = very unfamiliar). AI Tool Use Frequency indicates the weekly usage frequency of AI search tools.

ID	Age/Gender	Professional Background	AI Tool Familiarity	AI Tool Use Frequency	Study Area Familiarity	Presence in City
P1	28/F	Architecture	5	3-4	3	1 year
P2	27/F	Biomaterial Design	3	3-4	2	3.5 months
P3	32/M	Computer Science	5	5	2	8 months
P4	28/M	Social Science	5	5	4	11 months
P5	20/M	Chemistry	2	1-2	3	1.5 years
P6	39/X	Healthcare	3	1-2	4	7.5 years
P7	24/M	Machine Learning	5	5	2	10 months
P8	55/F	Art	3	1-2	5	30 years
P9	25/F	Education	4	5	1	11 months
P10	28/NB	Art	1	0	2	9 months
P11	21/F	Urban Development	3	3-4	4	7 months
P12	39/F	Carpenter	5	5	3	9.5 years
P13	22/M	Pure Mathematics	2	1-2	2	22 years
P14	45/F	Healthcare	4	3-4	5	36 years
P15	47/F	Dance	1	0	5	10 years
P16	57/F	Art	3	1-2	4	27 years
P17	29/M	Social Science	4	5	2	11 months
P18	24/M	Cell Biology	3	3-4	4	3 years
P19	35/M	Accessibility	4	5	2	9 months
P20	24/F	Tourism	2	1-2	1	2 weeks
P21	31/M	Material Science	1	0	4	11 years
P22	24/F	Business	4	5	3	3.5 years
P23	28/NB	Game Design	2	1-2	3	8 months

### 3.4 Data Analysis

To answer our RQs, we employed a re-existive thematic analysis (RTA) approach [6] to explore how participants interacted with UrbanSearch and reflected on their experiences. RTA emphasizes the researchers' active role in identifying, developing, and interpreting patterns of meaning across a dataset. Following this approach, the first and second authors led the coding process. We engaged iteratively in familiarizing ourselves with the data, generating initial codes, and developing themes that captured common behaviors and perspectives. Preliminary codes and insights were then discussed within the research team, where differing perspectives were compared and merged to lock in the final themes.

For the purpose of triangulation [1], we analyzed three data sources: (1) probe logs (including text and image inputs, probe responses), (2) screen recordings of participants' interactions, and (3) transcripts from post-session interviews. First, the first and second authors reviewed interview transcripts, probe logs, and screen recordings to build familiarity with the data. During this stage, we noticed a recurring pattern: participants often began with questions they already knew the answers to, or verified externally, as a way of testing UrbanSearch's capabilities. For example, one participant input the query, "Where am I?" despite already knowing

the starting location of the walk. We labeled such instances across all participants.

Beyond initial capability-testing queries, participants engaged in various topic-based interactions, where they genuinely sought answers, unfolding over multiple conversational turns. We grouped queries according to their thematic focus as topic-based interaction threads. To understand participants' information needs, we coded the topic of each interaction thread. Informed by prior research on everyday information needs [23, 26, 46], we used existing topic categories (e.g., Food and Drink) as deductive codes where applicable; when our data did not fit existing categories (e.g., interaction threads about local species), we developed inductive codes to capture these topics (e.g., Animal and Plant). In addition, to obtain more detailed insights for the queries by understanding what the participants were trying to accomplish with each query, we coded the task participants were performing about the topic with the given query. For example, for the topic Events & Activities, the query "Why people love heavy metal music festivals here?" shows the task Open-Ended Interpretation, whereas the query "Which day has the best show?" shows the task Comparison. We first identified tasks inductively using query data from five participants and collaboratively developed task categories through iterative team discussions. We then applied these categories to all queries to ensure consistent coding.

To analyze how information-seeking interactions unfolded within topic-based interaction threads, we borrowed the vocabulary from the OODA model [64] as prior codes, which have been widely used in HCI research [29]. We initially defined observe for queries to ask about their visual surroundings (e.g., Why doesn't the statue have a head?), decide for queries comparing options or applying criteria (e.g., which is cheaper?) and act for queries asked while taking an action (e.g., Am I on the right path?). We also coded screen recordings to label actions taken beyond UrbanSearch (e.g., opening Google Maps, completing a purchase). We excluded 14 queries that served only greetings (e.g., hello) or emotional expressions (e.g., cool!). We iteratively refined code boundaries through discussion to ensure they resonated with the data.

Next, the first and second authors independently conducted two cycles of open coding on the interview transcripts to better understand participants' behaviors and reactions on their experiences. In the first cycle, we used in vivo coding to stay close to participants' own language. In the second cycle, we applied pattern coding to identify conceptual groupings across participants. After the coding, the research team conducted iterative discussions to merge codes, compare interpretations, and refine candidate themes. Finally, the team revisited transcripts, logs, and recordings to check the coherence of the themes, and defined the scope and narrative arc of each, grounding them in illustrative quotes and log excerpts.

## 4 Findings

Participants on average engaged with UrbanSearch for 36.4 minutes (SD = 5.8). We detail our findings from these interactions in this section. After a high-level overview of the study sessions, we present our findings around the two RQs: the findings present values of using MCS tools and interaction dynamics with UrbanSearch (RQ1) and participants' expectations and concerns about using MCS technology in the outdoor urban space (RQ2).

### 4.1 Overview

Thirteen participants began the walking session with a specific initial goal that set the context for the individual sessions (e.g., going to a sunset spot; see Table 4 in Appendix A), while the remaining ten participants started with exploratory strolling. At the beginning, most participants (n = 19, 82.6%) engaged in a brief initial trust-building process with UrbanSearch through entering queries (M = 1.3 queries, SD = 0.5) with already known answers (detailed observations in Appendix D). Participants submitted 26.2 queries on average (SD = 4.1) and 19.0 camera-view images on average (SD = 4.3), resulting in a total of 602 queries and 451 camera-view images across all sessions. Participants engaged with UrbanSearch through an average of 10.5 topic-based interaction threads (SD = 2.1) shaped by the dynamic urban environment, with every participant engaging in multiple interaction threads. Participants' interaction threads spanned diverse topics (Table 2). On average, UrbanSearch responded to user queries with a latency of 2.8 seconds (SD = 1.9) and provided correct responses for user queries in 94.7% of the cases (excluding queries for which UrbanSearch did not provide a verifiable answer (e.g., I can't tell.)); details of the evaluation process and results in Appendix C)

### 4.2 Value of Using MCS Tools in Urban Space for Information Seeking

Participants' interactions with UrbanSearch and their reactions demonstrated the value of MCS tools for everyday information seeking in urban spaces, including requiring low effort in forming queries while providing highly relevant responses, and functioning as a central information gateway.

**4.2.1 Low-Effort Query Formulation for Responses Well-Situated in Shared Context.** The shared context (P1) between UrbanSearch and the participants, shaped by the probe's spatiotemporal awareness, allowed participants to ask questions conveniently and efficiently without providing the whole context. Specifically, participants frequently used brief time-space references, such as now to refer to the present time, here or nearby to indicate their location, and it or this to refer to an object in the scene they were looking at (captured image). Many participants (n = 16, 69.6%) mentioned the MCS tool enabled them to ask about environmental observations they usually ignored or bypassed previously. P16, who had lived in the city for 27 years, explained: I always wondered why this lamp on this street is so tall whenever I passed by. I never asked because it felt difficult to frame the question for a web search. This observation finally got an answer today. In another example, P7 saw a pond, took a picture, and asked Can I fish here? The participant noted that without MCS, he would have had to find and type the pond's name to conduct a web search with the whole sentence fishing regulation at [pond] in [city]. Together, MCS tools reduce two types of cognitive effort while walking: the conceptual load of having to translate an observation into searchable text (e.g., knowing what to call an extremely tall lamp and the pond's name) and the manual effort of forming (typing or speaking) a fully contextualized query. As a result, participants perceived MCS as sharing their immediate spatiotemporal context, which bridged them with the physical surroundings through situated conversations rather than environment-detached search tasks. We believe this reduced cognitive load lowers the barrier for participants to initiate information seeking on the move.

Additionally, many participants (n = 17, 73.9%) reported UrbanSearch's responses situated within the shared spatiotemporal context, which made the responses feel highly relevant. Beyond using participants' location and time as filters for decision-making (e.g., whether places were nearby and open), more than half of participants (n = 13, 56.5%) noted that this spatiotemporal awareness also made the system's interpretations more meaningful and tailored to their immediate environment. For example, P8 searched: Why do they grow grass on the roof? The probe tied its answers to the local ecosystem, referencing the urban climate and summer season. During the interview, P8 also searched the same question with a web search and ChatGPT. Both returned generic explanations about the benefits of growing grass on roofs, answers that apply broadly to any building, without referencing the participant's city or the specific summer season. As a result, the responses felt less relevant. P8 commented: When I asked this question, I was actually wondering if it had something to do with my location and season. Instead of giving me a generic answer that could apply anywhere and anytime, the response (from UrbanSearch) mentioned the local

Table 2: Topic distribution of interaction threads with descriptions and examples. Categories marked with \* are newly identified in this study with respect to the prior literature of everyday information needs [23, 26, 46]. # Threads represents the number of interaction threads related to the topic, and # Particip. indicates the number of participants who searched about the topic, respectively; the values in parentheses indicate the percentage.

Topic	Description	Example Queries	# Threads	# Particip.
Events & Activities	Specific local activities or events related to people and related information [23, 26].	Where can I pick up the wristband before Museum Night? ; What sports event is going on in the stadium?	47 (19.6%)	19 (82.6%)
Landmarks & Places*	Surrounding architecture, places of interest, infrastructure, and public arts.	How to visit the tallest building in the city? ; Who designed this building?	44 (18.4%)	20 (87.0%)
Food & Drink	Places to eat and drink, types of food served, dietary options, and dining-related information [46].	What is the best vegan lunch buffet nearby? ; How long do people usually wait in line for this place on Saturday?	42 (17.3%)	17 (73.9%)
Online Shopping	Retail stores, product availability, and shopping-related information [46].	Where can I buy second-hand skirts? ; Is there any other shop sell sun covers nearby?	32 (13.3%)	14 (60.9%)
Animals & Plants*	Local species and ecological significance.	Why do they grow grass on the roof? ; Is this bird a local species?	28 (11.6%)	16 (69.6%)
Social Norms & Pop Culture	Cultural behaviors, social etiquette, or pop culture references [26].	Do I need to pay tips? ; Is graffiti culture big here?	25 (10.4%)	13 (56.5%)
Regulation & Operation	Local rules, regulations, or operational details [26].	Can I fish here? ; How to complain about the water quality of the pond?	19 (7.8%)	8 (34.8%)
Environmental Conditions	Conditions of the physical environment include weather, traffic, and safety [26].	Is it safe to walk alone in this area? ; Is it going to rain soon?	4 (1.7%)	4 (17.4%)

ecosystem and explained it in relation to summer. That made me feel the response was highly relevant.

4.2.2 Functioning as a Central Information Gateway. Additionally, participants appreciated that UrbanSearch functioned as a central information gateway while walking, which allowed them to avoid frequently switching between multiple apps. Interaction threads ranged from 1 to 7 conversational turns ( $M = 2.4$ ,  $SD = 1.8$ ), and participants often asked multiple questions within a single topic that required multiple apps without an MCS tool. For example, in one interaction thread on the topic Animals & Plants, P3 noticed a pine cone, took a photo, asked whether it was edible, if it was a local species, and then asked where he could see the oldest pine tree in the city. He explained that such sequenced questions would normally require moving across several search engines and apps (Google Lens, web search, Maps), whereas the MCS allowed him to follow the entire line of inquiry within one interface and receive the aggregated information directly. In this way, the MCS functioned as a central information gateway that supported convenient information seeking while walking, since it lowered the cognitive effort needed to coordinate multiple apps and locate relevant information while navigating the environment.

### 4.3 Dissecting Interaction Threads: Observe, Decide, and Act

The topic-based interaction threads reflected the structure of the Observe-Decide-Act flow (example in Figure 3); participant notices something in their surroundings and asks about it (Observe), continues with queries that concretize a plan (Decide), and carries out the plan (Act). Shaped by the dynamic urban environment, participants' interaction threads were often fragmented and overlapped across stages of the Observe-Decide-Act flow, and only a few interaction

threads spanned the complete flow (Figure 9 in Appendix E for participant-level details).

We observed that many interaction threads ended at the Observe stage (Figure 4), often because their attention was drawn to new stimuli in the dynamic environment. For instance, P8 was asking about the relocation history of a zoo when she noticed a Van Gogh event poster and immediately dropped the previous interaction thread and shifted to a new topic. Also, many interaction threads ended at the Decide stage, we hypothesize that participants either decided not to take specific actions based on the decided plan or that they decided to table it for later due to the limited study duration. Additionally, although participants generally carried out actions in the Act stage outside the probe (e.g., going to a park following the Google Map, purchasing blueberries at a market), there were still some cases in which they used UrbanSearch during the Act stage. To provide a clearer view of what participants asked within each stage, Table 3 summarizes the types of query tasks observed across the Observe, Decide, and Act stages, along with their frequencies, descriptions, and example queries.

4.3.1 Observe. In the Observe stage, participants asked about objects and phenomena in their physical surroundings and sought to understand where they were and how they related spatially to the environment. Participants' queries fell into three task types: (1) Open-ended Interpretation, (2) Factual Information Retrieval, and (3) Spatial Orientation (Table 3 top segment).

Interestingly, participants asked many interpretive and reflective questions, such as "Why do people put stickers on the water pipe?" (P10) and "What do you think about the statue?" (P21). These interpretive queries reflected a desire not only to identify objects and retrieve factual information (e.g., designer, date, size), but also to explore their contextual meaning or social significance.

Figure 3: An illustration of an interaction thread that covers the full Observe Decide Act flow.

Figure 4: Flow of all participants' interaction threads through the Observe Decide Act stages. This diagram reflects patterns observed during the study's probe sessions and illustrates how participants transitioned across stages. No Action Taken means people only took actions outside of the UrbanSearch.

Table 3: Query tasks within Observe, Decide, and Act stage, with descriptions, query example, and number of queries within each stage, the values in parentheses indicate the percentage.

Stage	Task	Description	Example Query	# Query
Observe	Open-Ended Interpretation	Seeking interpretations or explanations for observed phenomena.	Why does this bench have a metal part in between?	193 (48.1%)
	Factual Information Retrieval	Asking for concrete facts about objects, places, or phenomena in view.	Which year was this church built?	154 (38.4%)
	Spatial Orientation	Locating oneself or understanding surrounding spatial relations.	Am I at the city center now?	54 (13.5%)
Decide	Pragmatic Evaluation	checking practical information such as cost, permissions, or feasibility.	How can I get a fishing license?	56 (31.6%)
	Comparison	Comparing options based on constraints such as distance, price, or time.	Which park is closer to me and easier to go to?	51 (28.8%)
	Ideation	Exploring ideas or possibilities for activities or destinations.	What can I do nearby for some history-related activities?	45 (25.4%)
	Route Planning	Planning which direction or path to take.	Which route to the park has the most shade?	25 (14.1%)
Act	Navigation	Providing directional guidance or support for moving toward a destination.	Should I turn left at this corner or the next one over there?	6 (60.0%)
	Creative Support	Offering guidance for creative tasks such as photography.	How to take a good picture of this scene?	3 (30.0%)
	Content Generation	Producing written content to support participants' immediate needs.	Draft a proposal to send to the government about water quality.	1 (10.0%)

We observed that most Observe queries were accompanied by an image as a visual reference. However, participants often encountered ambiguity in how their visual references (captured images) were interpreted by the UrbanSearch. When UrbanSearch struggled to determine which part of an image they were referencing,

participants (n = 20, 87.0%) felt that multi-turn conversations for clarifying visual references were unnecessarily disruptive. In P20's case, UrbanSearch detected two road signs in the image and asked which

sign the participant was referring to. P20 found this interaction frustrating: I had already passed the sign, and the probe said it couldn't tell which one I meant because there were two in my photo. I had to clarify again. This back and forth was pretty annoying it felt like solving an outdated issue after I'd already moved on. Just tell me both signs, or let me highlight the one when I ask.

**4.3.2 Decide.** In the Decide stage, participants often began with a vague or generic idea and concretized it into an actionable plan. For example, P14 searched with a vague intent, Where can I buy some blueberries? The probe initially suggested several markets across the city. P14 then requested a price and distance comparison through multi-turns of query, and ultimately decided to visit a specific open-air market. In addition, when the participants had difficulty coming up with criteria for evaluating their potential actions, UrbanSearch often helped formulate specific criteria through suggestions. For instance, P8 was seeking a second-hand clothing shop and asked, Which one should I go to? and UrbanSearch responded with I recommend option [A]. It is only a five-minute walk and is closer than the other two alternatives. It is also currently offering a mid-season sale, where a pair of jeans costs only 5 euros, making it more affordable than option [B] and [C]., effectively suggesting the evaluation criteria distance and cost for their decision-making process. Overall, participants' queries fell into four task types: (1) Pragmatic Evaluation, (2) Comparison, (3) Ideation, and (4) Route Planning (Table 3 middle segment).

Route Planning appeared when participants asked how to move toward a place. These queries often emphasized personalized needs that extend beyond the capabilities of current map applications. For instance, P12 asked, Can you tell me the route to [place], which way is most covered from the sun?

**4.3.3 Act.** In the Act stage, participants executed the action, including setting navigation routes in Google Maps, purchasing tickets or making reservations via external links, buying items at fruit vendors, submitting feedback through civic platforms, and taking photographs using the phone's default camera app. Although participants often carried out these actions outside the UrbanSearch, either through other apps or in-person behavior, UrbanSearch remained relevant as participants continued to ask Act queries during the actions (P4, 5, 11, 12, 15). Queries during the Act stage fell into three task types: (1) Navigation, (2) Creative Support, and (3) Content Generation (Table 3 bottom segment).

Many Act queries involved Navigation tasks that complemented pre-installed Google Maps. For example, P4 used Google Maps but intermittently switched back to UrbanSearch, taking photos of surrounding landmarks and asking whether he was on the correct side of the road. Additionally, participants also asked for creative support from the probe (P5, 11). For instance, P5 requested actionable tips on photographic composition and lighting for the scene. Interestingly, we also observed a Content Generation task. P15 asked UrbanSearch to draft a water quality reclamation proposal that she intended to submit to a government platform. Overall, these behaviors suggest that participants expected the MCS tool to continue providing support even after they had entered the action phase.

## 4.4 User Expectations for MCS Tools

Through interaction with UrbanSearch, participants expressed their expectations for MCS design, including visualized responses based on spatial reasoning, providing transparent integration of multiple information sources, bridging social interactions between users, and enabling flexible switch between text and audio interactions.

**4.4.1 Visualized Responses Based on Spatial Reasoning.** We observed that some queries went beyond retrieving text information and required the system to simulate and reason with spatial and temporal conditions in the urban environment. These queries cut across observe, decide, and act stages, showing a desire for the MCS tool to align with their immediate spatiotemporal context. We identified three types of spatial reasoning queries, illustrated in Figure 5.

**Map-View Alignment.** Participants expected UrbanSearch to align the live camera view with map or street view data. Queries such as Am I on the right way to [place]? (P4) or Where does this road lead me to? (P17) required UrbanSearch to reason about the user's location and orientation in relation to mapped routes or destinations.

**Spatial Simulation.** Participants sought reasoning that accounted for temporal context, orientation, and physical environmental conditions. For example, P12 asked, Which path is most covered from the sun to [place]? expecting UrbanSearch to simulate shade conditions along alternative routes. Similarly, queries like What is the direction of the tree's shadow? (P2) reflected expectations that the system could reason about orientation and temporal context in relation to visible objects.

**Spatial Creativity.** Beyond practical navigation, participants asked UrbanSearch to reason with spatial layouts to support creative activities. Examples included How to take a good picture of this scene? (P5) and How to decorate this empty space? (P13), which required a spatial interpretation of foreground, background, and compositional elements.

Across these cases, participants felt that generic text responses were insufficient. They expected visualized responses that integrated temporal context and spatial geometry. P12 described: I thought it (UrbanSearch) would show a mini 3D modeling and the actual path for me on the screen, but it just gave me a paragraph of text. That wasn't very useful. The participant provided a sketch (Figure 6) of the desired response format, which should be visualized to clearly show the 3D model of the spatiotemporal context.

**4.4.2 Transparent Integration of Multiple Information Sources.** UrbanSearch typically returned a single aggregated response without showing how different sources were considered. In contrast, participants themselves were often simultaneously engaging with at least two sources of information: their own direct observations and the external content retrieved by UrbanSearch. We observed that UrbanSearch sometimes introduced information from external sources that conflicted with what participants saw, leading to confusion and hesitation in decision-making. For example, P1 asked about the current temperature and attached an image of a rainy street. UrbanSearch responded: Based on the current weather forecast for [area], it is cloudy, with a mild temperature around 16-18°C. This puzzled P1 because the answer appeared to ignore her visible context and relied solely on external retrieved weather data.

Figure 5: Examples of participant queries that required the system to perform spatial reasoning.

Figure 6: A sketch from P12 illustrating their desired MCS response: a visualized path simulation that models the surrounding environment.

To this end, participants expected an MCS tool to explicitly acknowledge and present the multiple information sources it drew from, rather than masking them in a single output without transparency. As P10 reflected, they wished the system would surface different points of view, so they could make their own interpretation without losing multiple points of view. Similarly, P13 emphasized: The probe should acknowledge the multiple sources of information so we can critically assess them.

**4.4.3 Bridge Social Interaction between Users.** Some participants (n = 8, 34.8%) viewed MCS as a catalyst for shared experiences among friends and families walking together. For example, P23 described friends walking together with the technology as an intelligent sounding board that enhances conversations by providing factual knowledge or extra opinions: My friends and I often have friendly debates on things we see while we are wandering in the city, like why this is here, what does it mean. I could imagine this (MCS) technology actively enhancing our conversations giving us facts or simply providing another angle of thought. That would be fun.

**4.4.4 Flexible Audio/Text Interaction Switching in Social Contexts** Most participants (n = 19, 82.6%) preferred the audio interaction (spoke their queries and listened to audio replies) because it occupied less visual attention than the text while walking (P6, 7, 10, 19). Nonetheless, participants dynamically switched between audio and text depending on situational conditions of the urban space, such as

noise and social cues. When walking into noisy areas like bustling intersections or crowded markets, many (n = 18, 78.3%) mentioned they intentionally switched to the text modality, as the speech-to-text could not work well. In contrast, several participants (n = 8, 34.8%) mentioned that they used typing when they felt speaking aloud was socially inappropriate, such as in quiet parks or when standing near a street vendor. As P5 said: While I am in a crowd, I felt a bit uncomfortable speaking aloud, so I will type and mute the audio for a while. This flexible switching between text and audio allowed participants to align their interactions with the immediate social and spatial dynamics around them, and they expressed a desire for MCS tools to retain this design.

#### 4.5 Concerns for the MCS Technology in Everyday Usage

Despite the demonstrated values of MCS tools and the additional technical potential that participants expected from more advanced systems, participants also expressed concerns about peripheral and social awareness, and intrusion into personal space.

**4.5.1 Peripheral and Social Awareness.** While the MCS technology has the potential to bridge the user and the surroundings (Section 4.2.1) and other users (Section 4.4.3) through shared context, as a mobile app, it inherits the fundamental issues around peripheral and social awareness.

While most participants (n = 21, 91.3%) appreciated that UrbanSearch promoted them to inquire and interact with their surroundings, more than half (n = 12, 52.2%) observed that focusing on a specific queried object (e.g., a building or a bird) and interacting with the mobile device still reduced their broader awareness of nearby traffic while walking, which raised safety concerns.

In addition, some participants (n = 6, 26.1%) expressed concerns that using MCS tools might erode social interactions among people collocated in a shared public space. P14 was worried that relying too much on seeking answers from a mobile device might mean missing treasure moments of communicating with other people: One time I saw a person fishing, and I asked him about the process to get a license, which was a nice conversation, and I made a new friend. I guess I will get the same answer from the probe very fast and conveniently, but if I only use it, I would miss this chance to interact

with a real human. This reflection highlighted a concern that, in making information seeking efficient, the MCS technology might also make people less open to the serendipitous social encounters that enrich our urban life.

**4.5.2 Intrusion Into Personal Space.** While all participants were not concerned about sharing location data, which is pretty common practice in today's smartphone usage (P4), they expressed concerns about the capture and use of camera view data. All participants wanted to maintain control over the camera view capturing, rather than allowing continuous or automatic recording. As P23 explained: Although I like the camera-on design, which allows me to see and capture the surroundings conveniently, I would not let the tool automatically record all camera views. I want to control what I mean to ask and protect some privacy in this way. In one case, P10 realized during screen-recording review that their ID card had been accidentally captured in the camera view. This discovery left them unsettled: That is why I need to intentionally capture the photo and be prepared. If everything were automatically recorded, it could be risky. Beyond concerns about their own privacy, over half of the participants (n = 14, 53.8%) were also wary of unintentionally capturing other people's faces or behaviors without consent, potentially violating social boundaries.

This concern for personal space extended to how participants wanted the MCS tool to present its identity. Participants with more exposure to AI search tools (P1-4, 7, 9, 14, 17, 19, 22) preferred a more anthropomorphic design of the MCS tool: human-like voice and persona, or even playful role-play. For instance, P22, a regular user of AI tools, asked UrbanSearch to act as a bestie to enjoy the companionship of the tool during information searching. The rest of our participants, however, found human-like qualities unsettling, describing them as intrusive or even creepy. P15, who had no prior experience with AI tools, explained: While I am walking outside, it feels a bit scary if the system feels like a human. It feels like being stalked by someone. I would keep the current design, so it is a computer assistant with a robotic voice to me rather than a creepy stalker who sees what I see, knows where I am, and talks to me.

## 5 Discussion

Based on the findings, we discuss (1) potential everyday usage scenarios of MCS technology, (2) mitigating the risks of MCS in everyday use, and (3) design implications for MCS systems in urban spaces.

### 5.1 Potential Everyday Usage Scenarios of MCS Technology

Our technology probe sessions led participants to recognize the long-term potential of MCS tools in their everyday engagement with urban spaces. We found that MCS tools would support user needs across the stages of the Observe-Decide-Act flow that current web search or navigation tools alone cannot fully support. Three key future use scenarios surfaced.

First, participants envisioned MCS tools as enablers of environmental learning in everyday urban life, which is a crucial foundation of place-attachment [9] and lifelong learning [10]. Prior work notes that people often bypass opportunities for environmental learning, as it typically occurs as a secondary activity during commuting in

the city [18]. As shown in Section 4.2.1, MCS reduces the barriers to inquiring about environmental observations. Moreover, MCS functions as an information gateway and provides spatiotemporally relevant responses that connect the physical environment, the user, and the sea of information across the web and applications. Thus, MCS promotes the initiation and process of environmental learning.

Second, participants envisioned MCS tools as supporting the in-situ decisions that shape everyday urban life. Day-to-day activities involve numerous small, situational decisions (e.g., finding an ice cream truck nearby or visiting a sunset spot) that highly depend on the immediate spatiotemporal context (e.g., weather, time of the day, distance, personal energy, etc). In many cases, participants did not begin with a concrete goal (e.g., see sunset at a specific park 10 minutes away by walk) but rather with a vague intention (e.g., see sunset) that became more concrete through a multi-turn conversational search. As shown in Section 4.3.2, participants appreciated moments when MCS proactively compared options using real-world criteria such as distance and weather, offering decision support during movement.

Third, participants envisioned using MCS tools for personalized route planning and navigation as a complementary alternative to conventional map applications like Google Maps. When planning routes, participants appreciated how MCS tools could support personalized criteria, such as shaded paths or routes that avoid a specific block, that go beyond the heuristics like shortest or fastest used in conventional map apps. During movement toward a destination, participants acknowledged the usefulness of traditional navigation tools, especially the clear visual guidance provided by recent AR features like Live View. However, they also found value in MCS tools for resolving ambiguity in real-world settings through conversational visual queries. For instance, MCS can use the real-world references to help confirm whether they were on the correct path when routes were close together on the map (e.g., follow the trail with purple flowers, not the one with stones.), or for clarifying subtle turns in dense urban blocks where standard instructions or AR overlays might be confusing (e.g., turn right at the red mailbox.).

### 5.2 Mitigating the Risks of MCS in Everyday Use

Our findings (Section 4.5) identified concerns regarding MCS tools, including peripheral and social awareness, and intrusion into personal space. This section seeks potential methods of remedying these concerns.

#### 5.2.1 Mitigating the Concerns of Physical and Social Awareness.

Participants raised safety concerns due to the induced peripheral awareness while mobile phone usage (Section 4.5.1), a risk inherited from general mobile device use in public spaces [37] and walking [55]. Prior work has proposed using device cameras to detect nearby vehicles and warn distracted pedestrians [68] but continuous access to visual data raises privacy concerns for both users and bystanders. An alternative is to use GPS data to sense local traffic density and present spatial-awareness reminders on the interface, such as You are approaching a crosswalk on a busy street. Watch out for the traffic light and be careful of oncoming traffic.

Our findings (Section 4.5.1) reaffirm prior work [34] showing that using mobile devices while walking decreases users' attentiveness to other people. To realize participants' expectations of bridging social interactions with co-walkers (Section 4.4.3) and with others in shared public spaces, future systems may enable group chats between the users, with not only groups formed manually to include the co-walkers, but also groups formed with other people based on their proximity [14]. In the group chats, the system would engage in multi-way conversations, coordinating and guiding information sharing to reinforce the shared contexts between the users, not just between the users and the system, and thereby promote real-world social connections between people.

**5.2.2 Mitigating the Risk of Intruding on Personal Space.** Resonating with the literature [31, 41], our finding (Section 4.5.2) shows that the camera use in MCS raised data privacy concerns when recording in public spaces. To mitigate data privacy risks, MCS tools should prioritize intentional user control over what visual context is captured and used by the system, rather than defaulting to convenience-oriented automatic recording, which some emerging systems support [28]. In addition, future systems could leverage existing approaches (e.g., MediaPipe [54], text redaction [6, 11]) for automatically blurring people's faces and sensitive text information before an image is uploaded to an external server.

Beyond data privacy, our finding (Section 4.5.2) echoes the literature [3, 75], system anthropomorphism might cause negative feelings of being observed or socially monitored. While system anthropomorphism may enrich user experience and is a growing trend [22, 60, 76, 80, 98], some participants described human-like behaviors as provoking feelings of being stalked in physical space, especially when tied to location awareness. To mitigate this concern, an MCS system could offer different levels of human-likeness in its assistant identity (e.g., human, animal, or robot) through variations in speaking style and avatar appearance, softly nudging the users to choose the level of system anthropomorphism they feel most comfortable with.

### 5.3 Design Implications

Based on our findings, we propose four design implications (DIs) for future MCS systems in urban spaces.

**DI1: Support Compact Multi-Turn Conversational Search for Urban Mobile Contexts.** While multi-turn conversations helped participants clarify and extend their intent [58], they expressed frustration with clarification turns that only served confirmation purposes. In particular, participants found it distracting and time-consuming when the system asked clarification questions due to ambiguity in referring expressions, such as confirming which building they were asking about when multiple buildings were present. Although this seems to challenge the standard question-confirmation-feedback-answer mode [48, 82] in conventional conversational search systems, it instead highlights the prioritization of information immediacy in dynamic urban spaces. Future systems could first detect ambiguity and attempt to resolve it as much as possible, down to ideally couple possibilities. Then, instead of asking for additional conversation turns for disambiguation, it could present responses encompassing the possible interpretations to the user. The system should resort to asking for disambiguation only in cases

Figure 7: Proposed design strategies for compact multi-turn conversation: (a) allow the user to pinpoint the exact visual reference; (b) provide possible candidate answers directly.

where ambiguity persists beyond just a couple of possibilities that can be captured in a single response.

**Technical Considerations:** In MCS, the main challenges lie in detecting and addressing ambiguity in both text queries (e.g., the notion of 'nearby' in nearby supermarkets) and visual queries (e.g., this building in the context of an image with multiple buildings). To detect and address ambiguity in text queries, future work can utilize ambiguity detection and resolution methods (e.g., rule-based detection and widget-based resolution techniques [71], supervised neural network-based disambiguation [93], LLM-based ambiguity based on conversational context [47]). To detect and address ambiguity in visual queries, we propose two design strategies (see Figure 7). First, the system could enable users to intentionally clarify the referring object during query input, for example, by zooming the camera view, highlighting a region of the image, or pointing at the object with a finger in the frame. Zooming or highlighting would provide the system with a region of interest for object detection. The system can detect the fingertip and infer the pointing direction using hand-gesture detection approaches [33]. Although effective, this approach requires additional user effort to proactively disambiguate the query. Second, similar to ambiguity handling in pure text queries, the system could detect ambiguity in images [94, 96] and provide multi-candidate responses when the number of plausible targets is small (ideally two or three).

**DI2: Support Decision-Making Through Reasoning Over Multi-Source Information Hierarchy.** Handling conflicting information sources is a challenge in conversational search [50], and urban information environments often present inconsistencies between on-site observations and external data sources. Prior work suggests that

systems should surface multiple evidence sources and present conflicting viewpoints to users [50, 85]. However, for MCS in dynamic urban spaces, simply listing or summarizing multiple information sources might confuse users during real-world decision-making. Thus, future systems should support decision-making by reasoning over multi-source information and explaining which information might be more reliable when conflicts arise.

**Technical Considerations:** To realize this, the key challenges lie in detecting conflicting information sources and prioritizing the most relevant source for the response. The future system can adopt linguistic rule-based contradiction detection in natural language [25] and LLMs that can identify factual discrepancies and divergent opinions in text [89]. When such conflicts occur, we propose an information hierarchy model that ranks information sources based on their spatiotemporal relevance. The hierarchy has two core dimensions. First, spatial relevance, which prioritizes information sources based on their spatial proximity to the user; for example, city-level news is less relevant than neighborhood news where the user is standing. The second is temporal relevance, which prioritizes more up-to-date information sources; for example, a real-time camera view is more relevant than a weather forecast that might be delayed. In extreme cases, spatiotemporal relevance alone may not fully resolve conflicts between sources. For instance, an old road sign captured in the camera view (high spatial relevance) may contradict real-time map data (high temporal relevance). In such situations, the system should incorporate an additional reasoning layer that considers the query task and assigns priority accordingly. For geospatial tasks such as route planning or navigation, the system should prioritize real-time GPS and traffic data over static sources such as street-view imagery, which may be outdated. The interface design should reflect this information hierarchy when presenting results. When information sources provide consistent answers, the system can offer a single summarized response. When conflicts arise, the system should highlight the answer supported by the prioritized information source first with bold text, while also presenting information from conflicting sources using a distinguishable text color or visual indicator. Furthermore, the system should place an icon beside the response to prompt users to view the full list of sources ranked by relevance, as well as the underlying reasoning for the prioritization, which would guide the user toward deeper inspection for conflicting sources, thereby supporting informed decision-making.

**DI3: Provide Visualized Responses through World Modeling and Spatial Reasoning.** Prior work [52] identifies key qualities for conversational search systems, including managing dialogue history, summarizing responses, adapting to evolving user intent, and integrating world knowledge. Extending on this, our findings showed that participants did not merely want text information retrieval but also expected the MCS system to provide more visualized responses through simulating and reasoning with their specific spatiotemporal context (Section 4.4.1).

**Technical Considerations:** To realize this vision, key challenges lie in world modeling [51] and spatial reasoning [10]. World modeling requires not only reconstructing the 3D environment but also inferring its underlying physical state and dynamics how objects, people, lighting, and weather change over time, as well as how the

user's perspective relates to broader, unseen urban surroundings which can be partially addressed by LiDAR-based scene geometry [99] and VLMs capable of object goal navigation [69]. The future system can leverage the 3D visual grounding datasets [57] about natural language descriptions of spatial relationships between objects at the city-level as a geographical grounding benchmark for the system's global, allocentric spatial awareness. Spatial reasoning builds upon this world model, requiring the ability to ground user queries in the modeled 3D environment, infer spatial relations from an egocentric perspective, and simulate what the user might see or encounter next in the urban space. The future system can leverage the dataset for vision-and-language navigation [21] to train the system's egocentric, situated spatial reasoning and navigation capabilities. As such, the MCS system can bridge cyberspace and the physical world, acting not just as information retrievers but as situated, spatially aware assistants. Yet, the system design must rigorously control the latency introduced by spatial computing to avoid harming the immediacy of information delivery.

**DI4: Facilitate Seamless Integration of External Tools for Taking Actions.** Future systems should interoperate with external tools to better support action-taking. While participants often relied on third-party apps such as Google Maps or event websites to carry out their intended actions, the MCS system still often served as the central gateway surfacing place names, event titles, or relevant links. However, participants frequently described a disjointed experience when transitioning from UrbanSearch to external apps, often needing to manually set up the search context again. This interruption not only broke the flow of interaction but also imposed additional effort at a moment when their attention was already divided in the urban mobile context.

**Technical Considerations:** To address this, future systems should support interoperability with external apps on the smartphones [3]. The future system can leverage LLM-driven user interface automation [86, 92] that automates operations such as launching target apps, entering text for searching, clicking, and scrolling. The future system first needs to detect whether the system response requires information transfer to an external app, classifying each (query, response) pair into a targeted external app (e.g., map, booking app) or none. It can be done by using simple rule-based patterns (e.g., detecting phrases like "how do I get to..." or "book a ticket") or prompting LLMs [86]. If needed, the system can use an LLM-based user interface planner to construct an app-automation plan, specifying which app to open, what information to transfer, and which interface steps to execute [92]. Although full automation is technically feasible, maintaining user agency and ensuring correct information transfer still requires explicit user verification. To support user awareness during cross-app communication, the system interface can highlight the transferred information in the response (e.g., the place name) and, in the newly opened app, require confirmation at key steps (e.g., tapping the search button).

## 6 Limitations and Future Work

**Limitations of Study Design.** As an early work in the domain, our study included several limitations while contributing an initial set of understanding around user interactions with MCS technology. To begin with, we conducted the study with a single session of a

30-minute technology probing session. Although this design was sufficient for making observations of interaction patterns beyond the trust-building process and understanding the values of MCS tools, and the expectations and concerns about the technology, we believe that there are interaction patterns that would surface in longer repeated use of the tool, which would require longitudinal studies. Second, our probing sessions were conducted in a single European city. While we believe our findings are generalizable, future work should perform studies in other culturally different cities around the world with similar diversification of participants. Third, while participants were free to perform technology probing either in the context of exploratory strolling or in the context of an initial goal, the context of the technology probing may affect interactions and suggest controlled studies to further investigate the effects of such context. Finally, we note that although a researcher was present at a distance during the technology probing session (Section 3.3.2) to ensure participants' safety, this necessary measure could have nudged participants to use the tool more frequently and avoid socially inappropriate queries.

**Further Study & Analysis** Building on our results showing that MCS tools reduce effort for information seeking by lowering the conceptual load of formulating search queries from observations and reducing the manual effort of contextualizing the queries, future work could further study the various ways in which MCS tools reduce user efforts and organize them into a taxonomy. In addition, while our findings report the accuracy of UrbanSearch's responses (Section 4) and explore how users calibrate trust in the tool accordingly (Appendix D), future research could further examine how users perceive correctness and how such perceptions shape their decision-making and reliance on the system.

**Beyond Everyday Information Search.** Several participants mentioned potential applications in travel scenarios. For example, we envisioned using an MCS tool as a traveling assistant in a foreign country, leveraging it to plan spontaneous nearby activities. We believe that traveling in a city may introduce different interaction priorities than those in everyday routines, such as overcoming language barriers and cultural differences. Hence, exploring MCS in other scenarios and specific contexts (e.g., travel) is a valuable next step for understanding how use cases affect real-time search behavior in urban spaces.

**Beyond the Use of Mobile Phone.** While smartphones will likely remain the mainstream device for conversational search in the near future, emerging hands-free wearables such as smart glasses may open new opportunities and challenges once they become more affordable and widely adopted [8, 73]. These devices could make conversational search more seamlessly integrated into urban movement, reducing the friction of holding and operating a phone. At the same time, they raise unique questions about social acceptability and privacy in public spaces. Future research should therefore examine how social norms and privacy aspects shape the dynamics of situated information seeking, including how users negotiate attention and control in urban spaces using different kinds of mobile devices beyond smartphones.

## 7 Conclusion

This work adds to the growing endeavor of understanding MCS by providing insights into how people engage with spatiotemporal-aware conversational search systems while walking in urban spaces. We developed a technology probe, UrbanSearch, and conducted a contextual inquiry with 23 participants, discovering that MCS tools provide two core values: (1) requiring low effort in forming queries while offering highly relevant responses, and (2) functioning as a central information gateway. The findings suggest that MCS support everyday user scenarios such as environmental learning, in-situ decision making, and personalized navigation. At the same time, we also revealed unmet needs for spatial reasoning, transparent integration of multi-source information, and concerns related to social awareness, peripheral attention, and intrusion into personal space. Our findings yielded four design implications on enabling compact multi-turn conversational search, ensuring transparency across information sources, advancing world modeling and spatial intelligence, and supporting interoperability with external tools. In conclusion, our work proposes that MCS technology brings additional value to traditional search paradigms, evolving beyond information retrieval toward contextually adaptive, socially sensitive, and spatially grounded companions, and opening new research agendas at the intersection of people, information seeking, and urban space.

## Acknowledgments

Dae Hyun Kim and Suyeon Seo were supported by the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2026 (Project Name: Development of Multimodal UX Evaluation Platform Technology for XR Spatial Responsive Content Optimization, Grant No.: RS-2024-00361757, Contribution Rate: 30%) and the 2025-1 Yonsei University Future-Leading Research Initiative (Grant No.: 2025-22-0156). Joongi Shin was supported by the Research Council of Finland (Subjective Functions: 357578) and the ERC Advanced Grant (101141916). Alan Chamberlain was supported by the Turing AI World Leading Researcher Fellowship in Somabotics: Creatively Embodying Artificial Intelligence (Grant No.: EP/Z534808/1) and AI UK: Creating an International Ecosystem for Responsible AI Research and Innovation (Grant No.: EP/Y009800/1). We used ChatGPT to perform grammar checks, phrase suggestions, and grammatical restructuring based on our own texts.

## References

- [1] Rafa Absar, Heather O'Brien, and Eric T. Webster. 2014. Exploring Social Context in Mobile Information Behavior. *Proceedings of the American Society for Information Science and Technology* 51, 1 (2014), 1-10. <https://doi.org/10.1002/meet.2014.14505101058>
- [2] Utku Günay Acer, Marc van den Broeck, Chulhong Min, Malleham Dasari, and Fahim Kawsar. 2022. The City as a Personal Assistant: Turning Urban Landmarks into Conversational Agents for Serving Hyper Local Information. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 40 (July 2022), 31 pages. <https://doi.org/10.1145/3534573>
- [3] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2015. Privacy and Human Behavior in the Age of Information. *Science* 347, 6221 (2015), 509-514. <https://doi.org/10.1126/science.aaa1465>
- [4] Denise E. Agosto and Sandra Hughes-Hassell. 2005. People, Places, and Questions: An Investigation of the Everyday Life Information-Seeking Behaviors

- of Urban Young Adults. *Library & Information Science Research* 27, 2 (2005), 141–163. <https://doi.org/10.1016/j.lisr.2005.01.002>
- [5] Taneesa S Agrawaal, Aarjav Chauhan, Carolina Nobre, and Robert Soden. 2024. What's the Rush?: Alternative Values in Navigation Technologies for Urban Placemaking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 501, 17 pages. <https://doi.org/10.1145/3613904.3642470>
- [6] Hadeer Ahmed, Issa Traore, Sherif Saad, and Mohammad Mamun. 2021. Automated Detection of Unstructured Context-Dependent Sensitive Information Using Deep Learning. *Internet of Things* 16 (2021), 100444.
- [7] Mutahar Ali, Arjun Arunasalam, and Habiba Farrukh. 2025. Understanding Users' Security and Privacy Concerns and Attitudes Towards Conversational AI Platforms. In *2025 IEEE Symposium on Security and Privacy (SP)*. Institute of Electrical and Electronics Engineers, New York, NY, USA, 298–316. <https://doi.org/10.1109/SP61157.2025.00241>
- [8] Mohammad Aliannejadi, Morgan Harvey, Luca Costa, Matthew Pointon, and Fabio Crestani. 2019. Understanding Mobile Search Task Relevance and User Behaviour in Context. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (Glasgow, Scotland UK) (CHIIR '19)*. Association for Computing Machinery, New York, NY, USA, 143–151. <https://doi.org/10.1145/3295750.3298923>
- [9] Irwin Altman and Setha M Low. 2012. *Place Attachment*. Vol. 12. Springer Science & Business Media, New York, NY, USA and London, UK.
- [10] Nicole M. Ardoin and Joe E. Heimlich. 2021. Environmental Learning in Everyday Life: Foundations of Meaning and a Context for Change. *Environmental Education Research* 27, 12 (2021), 1681–1699. <https://doi.org/10.1080/13504622.2021.1992354>
- [11] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. 2019. What is Wrong with Scene Text Recognition Model Comparisons? Dataset and Model Analysis. In *Proceedings of the IEEE/CVF international conference on computer vision*. Institute of Electrical and Electronics Engineers, New York, NY, USA, 4715–4723. <https://doi.org/10.1109/ICCV.2019.00481>
- [12] Hugh Beyer and Karen Holtzblatt. 1999. Contextual Design. *Interactions* 6, 1 (Jan. 1999), 32–42. <https://doi.org/10.1145/291224.291229>
- [13] Yiheng Bian, Yunpeng Song, Guiyu Ma, Rongrong Zhu, and Zhongmin Cai. 2025. DroidRetriever: An Autonomous Navigation and Information Integration System Facilitating Mobile Sensemaking. [arXiv:2505.03364 \[cs.HC\]](https://arxiv.org/abs/2505.03364) <https://arxiv.org/abs/2505.03364>
- [14] Erik W Black, Kelsey Mezzina, and Lindsay A Thompson. 2016. Anonymous Social Media Understanding the Content and Context of Yik Yak. *Computers in Human Behavior* 57 (2016), 17–22.
- [15] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, Karthik Padthe, Vasu Sharma, Hu Xu, Xiaoqing Ellen Tan, Megan Richards, Samuel Lavoie, Pietro Astol, Reyhane Askari Hemmat, Jun Chen, Kushal Tirumala, Rim Assouel, Mazda Moayeri, Arjang Talattof, Kamalika Chaudhuri, Zechun Liu, Xilun Chen, Quentin Garrido, Karen Ullrich, Aishwarya Agrawal, Kate Saenko, Asli Celikyilmaz, and Vikas Chandr. 2024. An Introduction to Vision-Language Modeling. [arXiv:2405.17247 \[cs.LG\]](https://arxiv.org/abs/2405.17247) <https://arxiv.org/abs/2405.17247>
- [16] Virginia Braun, Victoria Clarke, Nikki Hayeld, Louise Davey, and Elizabeth Jenkinson. 2022. *Doing Reflexive Thematic Analysis*. Springer International Publishing, Cham, 19–38. [https://doi.org/10.1007/978-3-031-13942-0\\_2](https://doi.org/10.1007/978-3-031-13942-0_2)
- [17] Sally Burford and Sora Park. 2014. The Impact of Mobile Tablet Devices on Human Information Behaviour. *Journal of Documentation* 70, 4 (07 2014), 622–639. <https://doi.org/10.1108/JD-09-2012-0123>
- [18] Runze Cai, Nuwan Janaka, Hyeoncheol 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 (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 631, 26 pages. <https://doi.org/10.1145/3706598.3713953>
- [19] Jessica R. Cauchard, Jane L. E. Kevin Y. Zhai, and James A. Landay. 2015. Drone & me: an exploration into natural human-drone interaction. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (Osaka, Japan) (UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 361–365. <https://doi.org/10.1145/2750858.2805823>
- [20] Joseph Chee Chang, Nathan Hahn, and Aniket Kittur. 2016. Supporting Mobile Sensemaking Through Intentionally Uncertain Highlighting. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. Association for Computing Machinery, New York, NY, USA, 61–68. <https://doi.org/10.1145/2984511.2984538>
- [21] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural Language Navigation and Spatial Reasoning in Visual Street Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, New York, NY, USA, 12538–12547*. <https://doi.org/10.1109/CVPR.2019.01282>
- [22] Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From Persona to Personalization: A Survey on Role-Playing Language Agents. *Transactions on Machine Learning Research* 2024 (2024), 50 pages. <https://openreview.net/forum?id=xrO70E8UIZ> Survey Certification.
- [23] Karen Church, Mauro Cherubini, and Nuria Oliver. 2014. A Large-Scale Study of Daily Information Needs Captured in Situ. *ACM Transactions on Computer-Human Interaction* 21, 2, Article 10 (Feb. 2014), 46 pages. <https://doi.org/10.1145/2552193>
- [24] Karen Church and Barry Smyth. 2009. Understanding the Intent Behind Mobile Information Needs. In *Proceedings of the 14th International Conference on Intelligent User Interfaces (UI '09)*. Association for Computing Machinery, New York, NY, USA, 247–256. <https://doi.org/10.1145/1502650.1502686>
- [25] Marie-Catherine De Marne e, Anna N Rarty, and Christopher D Manning. 2008. Finding Contradictions in Text. In *Proceedings of acl-08: Hlt*. Association for Computational Linguistics, Columbus, Ohio, USA, 1039–1047.
- [26] David Dearman, Melanie Kellar, and Khai N. Truong. 2008. An Examination of Daily Information Needs and Sharing Opportunities. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW '08)*. Association for Computing Machinery, New York, NY, USA, 679–688. <https://doi.org/10.1145/1460563.1460668>
- [27] Guy Debord. 1955. *Introduction to a Critique of Urban Geography*. Les Lèvres Nues, Antwerp, Belgium.
- [28] Google DeepMind and Google AI. 2025. Gemini 2.5 Pro [Large multimodal language model]. Retrieved Jan 18, 2026 from <https://ai.google.dev/gemini-api/docs/models#gemini-2.5-pro>
- [29] Huseyin Dogan, Stephen Gi, and Renee Barsoum. 2024. User Experience Research: Point of View Playbook. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 537, 7 pages. <https://doi.org/10.1145/3613905.3637136>
- [30] Marian Dörk, Sheelagh Carpendale, and Carey Williamson. 2011. The Information Flaneur: A Fresh Look at Information Seeking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 1215–1224. <https://doi.org/10.1145/1978942.1979124>
- [31] Marco Furini, Silvia Mirri, Manuela Montangero, and Catia Prandi. 2020. Privacy Perception when Using Smartphone Applications. *Mobile Networks and Applications* 25, 3 (2020), 1055–1061. <https://doi.org/10.1007/s11036-020-01529-z>
- [32] Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G. Karahalios. 2015. DataTone: Managing Ambiguity in Natural Language Interfaces for Data Visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology (Charlotte, NC, USA) (UIST '15)*. Association for Computing Machinery, New York, NY, USA, 489–500. <https://doi.org/10.1145/2807442.2807478>
- [33] Google. 2025. Google Maps. Retrieved September 6, 2025 from <https://maps.google.com>
- [34] Google. 2025. Watch Google's Android XR Glasses Demo from I/O 2025. Retrieved August 23, 2025 from <https://blog.google/products/android/android-xr-glasses-demo-io-2025>
- [35] Hans W. Guesgen and Stephen Marsland. 2010. *Spatio-Temporal Reasoning and Context Awareness*. Springer US, Boston, MA, 609–634. [https://doi.org/10.1007/978-0-387-93808-0\\_23](https://doi.org/10.1007/978-0-387-93808-0_23)
- [36] Ece Gumusel, Kyrie Zhixuan Zhou, and Madelyn Rose Sanlippo. 2024. User Privacy Harms and Risks in Conversational AI: A Proposed Framework. [arXiv:2402.09716 \[cs.HC\]](https://arxiv.org/abs/2402.09716)
- [37] Morgan Harvey and Matthew Pointon. 2017. Searching on the Go: The Effects of Fragmented Attention on Mobile Web Search Tasks. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 155–164. <https://doi.org/10.1145/3077136.3080770>
- [38] Kristian Hentschel, Dejice Jacob, Jeremy Singer, and Matthew Chalmers. 2016. Supersensors: Raspberry Pi Devices for Smart Campus Infrastructure. In *2016 IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud)*. IEEE Computer Society, Vienna, Austria, 58–62. <https://doi.org/10.1109/FiCloud.2016.16>
- [39] Orland Hoerber, Morgan Harvey, Shaheed Ahmed Dewan Sagar, and Matthew Pointon. 2022. The Effects of Simulated Interruptions on Mobile Search Tasks. *Journal of the Association for Information Science and Technology* 73, 6 (2022), 777–796. <https://doi.org/10.1002/asi.24579>
- [40] Stewart Home et al 1996. *What Is Situationism?: A Reader*. AK Press, Edinburgh, Scotland and San Francisco, CA.
- [41] Roberto Hoyle, Robert Templeman, Steven Armes, Denise Anthony, David Crandall, and Apu Kapadia. 2014. Privacy Behaviors of Lifeloggers Using Wearable

- Cameras. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14). Association for Computing Machinery, New York, NY, USA, 571 582. <https://doi.org/10.1145/2632048.2632079>
- [42] Peng-Kai Hung, Janet Yi-Ching Huang, Rung-Huei Liang, and Stephan Wensveen. 2025. Generative AI as a Playful yet Offensive Tourist: Exploring Tensions Between Playful Features and Citizen Concerns in Designing Urban Play. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 205, 20 pages. <https://doi.org/10.1145/3706598.3713137>
- [43] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology Probes: Inspiring Design for and with Families. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03). Association for Computing Machinery, New York, NY, USA, 17 24. <https://doi.org/10.1145/642611.642616>
- [44] Jonas Ivarsson and Oskar Lindwall. 2023. Suspicious Minds: The Problem of Trust and Conversational Agents. *Computer Supported Cooperative Work (CSCW)* 32, 3 (2023), 545 571.
- [45] Jorge Joo-Nagata and Jorge Rodríguez-Becerra. 2025. Mobile Pedestrian Navigation, Mobile Augmented Reality, and Heritage Territorial Representation: Case Study in Santiago de Chile. *Applied Sciences* 15, 6 (2025), 19 pages. <https://doi.org/10.3390/app15062909>
- [46] Maryam Kamvar and Shumeet Baluja. 2006. A Large Scale Study of Wireless Search Behavior: Google Mobile Search. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06). Association for Computing Machinery, New York, NY, USA, 701 709. <https://doi.org/10.1145/1124772.1124877>
- [47] Aryan Keluskar, Amrita Bhattacharjee, and Huan Liu. 2024. Do LLMs Understand Ambiguity in Text? A Case Study in Open-World Question Answering. In 2024 IEEE International Conference on Big Data (BigData). IEEE, Institute of Electrical and Electronics Engineers (IEEE), Washington, DC, USA, 7485 7490.
- [48] Kimiya Keyvan and Jimmy Xiangji Huang. 2022. How to Approach Ambiguous Queries in Conversational Search: A Survey of Techniques, Approaches, Tools, and Challenges. *Comput. Surveys* 55, 6, Article 129 (Dec. 2022), 40 pages. <https://doi.org/10.1145/3534965>
- [49] Hannu Kukka, Vassilis Kostakos, Timo Ojala, Johanna Ylipulli, Tiina Suopajarvi, Marko Jurmu, and Simo Hosio. 2013. This Is Not Classi ed: Everyday Information Seeking and Encountering in Smart Urban Spaces. *Personal and Ubiquitous Computing* 17, 1 (Jan. 2013), 15 27. <https://doi.org/10.1007/s00779-011-0474-1>
- [50] Weronika Sąjewska, Damiano Spina, Johanne Trippas, and Krisztian Balog. 2024. Explainability for Transparent Conversational Information-Seeking. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24). Association for Computing Machinery, New York, NY, USA, 1040 1050. <https://doi.org/10.1145/3626772.3657768>
- [51] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. 2025. Aligning Cyber Space With Physical World: A Comprehensive Survey on Embodied AI. *IEEE/ASME Transactions on Mechatronics* 30, 6 (2025), 1 22. <https://doi.org/10.1109/TMECH.2025.3574943>
- [52] Zhanyi Liu, Zheng-Yu Niu, Jian-Yun Nie, Hua Wu, and Haifeng Wang. 2017. Conversation in IR: Its Role and Utility. In 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17), Vol. 222. Association for Computing Machinery, New York, NY, USA, 4 pages.
- [53] Irene Lopatovska, Katrina Rink, Ian Knight, Kieran Raines, Kevin Cosenza, Harriet Williams, Perachya Sorsche, David Hirsch, Qi Li, and Adrianna Martinez. 2019. Talk to Me: Exploring User Interactions with the Amazon Alexa. *Journal of Librarianship and Information Science* 51, 4 (2019), 984 997. <https://doi.org/10.1177/0961000618759414>
- [54] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A Framework for Building Perception Pipelines. *arXiv preprint arXiv:1906.08172* (2019), 9 pages.
- [55] Munir Makhmutov, Timur Asapov, and Joseph Alexander Brown. 2021. Safety Risks in Location-Based Augmented Reality Games. In Entertainment Computing ICEC 2021: 20th IFIP TC 14 International Conference, ICEC 2021, Coimbra, Portugal, November 2 5, 2021, Proceedings. Springer-Verlag, Berlin, Heidelberg, 457 464. [https://doi.org/10.1007/978-3-030-89394-1\\_39](https://doi.org/10.1007/978-3-030-89394-1_39)
- [56] Colin McFarlane. 2011. The City as Assemblage: Dwelling and Urban Space. *Environment and Planning D: Society and Space* 29, 4 (2011), 649 671. <https://doi.org/10.1068/d4710>
- [57] Taiki Miyanishi, Fumiya Kitamori, Shuhei Kurita, Jungdae Lee, Motoaki Kawanabe, and Nakamasa Inoue. 2023. CityRefer: Geography-Aware 3D Visual Grounding Dataset on City-scale Point Cloud Data. In Proceedings of the 37th International Conference on Neural Information Processing Systems (New Orleans, LA, USA) (NIPS '23). Curran Associates Inc., Red Hook, NY, USA, Article 3397, 13 pages.
- Fengran Mo, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Jian-Yun Nie. 2025. A Survey of Conversational Search. *ACM Transactions on Information Systems* 43, 6, Article 167 (Sept. 2025), 50 pages. <https://doi.org/10.1145/3759453>
- [59] Dujun Nie, Xianda Guo, Yiqun Duan, Ruijun Zhang, and Long Chen. 2025. Wmnav: Integrating Vision-Language Models into World Models for Object Goal Navigation. *arXiv preprint arXiv:2503.02247* (2025).
- [60] Ruowen Niu, Jiaxiang Hu, Siyu Peng, Caleb Chen Cao, Chengzhong Liu, Sirui Han, and Yike Guo. 2025. Scenario, Role, and Persona: A Scoping Review of Design Strategies for Socially Intelligent AI Agents. In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25). Association for Computing Machinery, New York, NY, USA, Article 470, 9 pages. <https://doi.org/10.1145/3706599.3719762>
- [61] Alice Olijie Odu and Madely du Preez. 2023. Exploring the Information-seeking Behaviour of Students at the Federal University of La a, Nigeria, who Use Mobile Technologies to Access Information. *Mousaion* 41, 4 (2023), 1 19. <https://doi.org/10.25159/2663-659X/12697>
- [62] Timo Ojala, Vassilis Kostakos, Hannu Kukka, Tommi Heikkinen, Tomas Linden, Marko Jurmu, Simo Hosio, Fabio Kruger, and Daniele Zanni. 2012. Multipurpose Interactive Public Displays in the Wild: Three Years Later. *Computer* 45, 5 (2012), 42 49. <https://doi.org/10.1109/MC.2012.115>
- [63] OpenAI. 2025. ChatGPT. Retrieved June 14, 2025 from <https://chat.openai.com/chat>
- [64] Frans PB Osinga. 2007. *Science, Strategy and War: The Strategic Theory of John Boyd*. Routledge.
- [65] Jie Qi, Suvodeep Mazumdar, and Ana C Vasconcelos. 2024. Understanding the Relationship between Urban Public Social Cohesion: A Systematic Review. *International Journal of Community Well-Being* 7, 2 (2024), 155 212. <https://doi.org/10.1007/s42413-024-00204-5>
- [66] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17). Association for Computing Machinery, New York, NY, USA, 117 126. <https://doi.org/10.1145/3020165.3020183>
- [67] Abu Rayhan and David Gross. 2025. SearchGPT: Revolutionizing Information Retrieval with Advanced Language Models. (2025), 10 pages.
- [68] Nikos A. Salingaros. 1999. Urban Space and Its Information Field. *Journal of Urban Design* 4, 1 (1999), 29 49. <https://doi.org/10.1080/13574809908724437>
- [69] Reijo Savolainen. 1995. Everyday Life Information Seeking: Approaching Information Seeking in the Context of Way of Life. *Library & Information Science Research* 17, 3 (1995), 259 294. [https://doi.org/10.1016/0740-8188\(95\)90048-9](https://doi.org/10.1016/0740-8188(95)90048-9)
- [70] ScienceDirect. 2025. Urban Spaces. Retrieved September 9, 2025 from <https://www.sciencedirect.com/topics/social-sciences/urban-spaces>
- [71] Vidya Setlur, Sarah E. Battersby, Melanie Tory, Rich Gossweiler, and Angel X. Chang. 2016. Eviza: A Natural Language Interface for Visual Analysis. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology (Tokyo, Japan) (UIST '16). Association for Computing Machinery, New York, NY, USA, 365 377. <https://doi.org/10.1145/2984511.2984588>
- [72] Saguna Shankar, Heather L O'Brien, and Rafa Absar. 2018. Rhythms of Everyday Life in Mobile Information Seeking: Re actions on a Photo-Diary Study. *Library Trends* 66, 4 (2018), 535 567.
- [73] Skift. 2025. Google's New Smart Glasses: Live Translation, Navigation. Retrieved August 23, 2025 from <https://skift.com/2025/05/21/googles-new-smart-glasses-for-travelers-live-translation-navigation/>
- [74] Timothy Sohn, Kevin A. Li, William G. Griswold, and James D. Hollan. 2008. A Diary Study of Mobile Information Needs. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08). Association for Computing Machinery, New York, NY, USA, 433 442. <https://doi.org/10.1145/1357054.1357125>
- [75] Avelie Stuart, Arosha K. Bandara, and Mark Levine. 2019. The Psychology of Privacy in the Age. *Social and Personality Psychology Compass* 13, 11 (2019), e12507. <https://doi.org/10.1111/spc3.12507>
- [76] Guangzhi Sun, Xiao Zhan, and Jose Such. 2024. Building Better AI Agents: A Provocation on the Utilisation of Persona in LLM-based Conversational Agents. In Proceedings of the 6th ACM Conference on Conversational User Interfaces (CUI '24). Association for Computing Machinery, New York, NY, USA, Article 35, 6 pages. <https://doi.org/10.1145/3640794.3665887>
- [77] Xin Sun, Rongjun Ma, Xiaochang Zhao, Zhuying Li, Janne Lindqvist, Abdallah El Ali, and Jos A. Bosch. 2024. Trusting the Search: Unraveling Human Trust in Health Information from Google and ChatGPT. *arXiv:2403.09987 [cs.HC]* <https://arxiv.org/abs/2403.09987>
- [78] Sakari Tamminen, Antti Oulasvirta, Kalle Toiskallio, and Anu Kankainen. 2004. Understanding Mobile Contexts. *Personal and Ubiquitous Computing* 8, 2 (2004), 135 143. <https://doi.org/10.1007/s00779-004-0263-1>
- [79] Jaime Teevan, Amy Karlson, Shahriyar Amini, A. J. Bernheim Brush, and John Krumm. 2011. Understanding the Importance of Location, Time, and People in Mobile Local Search Behavior. In Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '11). Association for Computing Machinery, New York, NY, USA, 77 80. <https://doi.org/10.1145/1958448.1958487>

- //doi.org/10.1145/2037373.2037386
- [80] Luke Thominet, Jacqueline Amorim, Kristine Acosta, and Vanessa K. Sohan. 2024. Role Play: Conversational Roles as a Framework for Reflexive Practice in AI-Assisted Qualitative Research. *Journal of Technical Writing and Communication* 54, 4 (2024), 396–418. <https://doi.org/10.1177/00472816241260044>
- [81] Veronica A. Thurmond. 2001. The Point of Triangulation. *Journal of Nursing Scholarship* 33, 3 (2001), 253–258. <https://doi.org/10.1111/j.1547-5069.2001.00253.x>
- [82] Svitlana Vakulenko, Kate Revoredo, Claudio Di Ciccio, and Maarten de Rijke. 2019. QRFA: A Data-Driven Model of Information-Seeking Dialogues. In *Advances in Information Retrieval*. Springer International Publishing, Cham, 541–557. [https://doi.org/10.1007/978-3-030-15712-8\\_35](https://doi.org/10.1007/978-3-030-15712-8_35)
- [83] Andrey Vakunov, Chuo-Ling Chang, Fan Zhang, George Sung, Matthias Grundmann, and Valentin Bazarevsky. 2020. Mediapipe Hands: On-Device Real-Time Hand Tracking. In *Workshop on Computer Vision for AR/VR*, Vol. 2. arXiv, Seattle, WA, USA, 5 pages.
- [84] Imre van Kraalingen and Simon Beames. 2024. Presence and (Dis)connectedness – The Influence of Smartphones Usage on Human–Nature and Human–Human Interactions in Outdoor Studies. *Frontiers in Education* Volume 9 - 2024 (2024), 12 pages. <https://doi.org/10.3389/feeduc.2024.1369591>
- [85] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles L. A. Clarke. 2017. Exploring Conversational Search With Humans, Assistants, and Wizards. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 2187–2193. <https://doi.org/10.1145/3027063.3053175>
- [86] Minh Duc Vu, Han Wang, Jieshan Chen, Zhuang Li, Shengdong Zhao, Zhenchang Xing, and Chunyang Chen. 2024. GPTVoiceTasker: Advancing Multi-step Mobile Task Efficiency Through Dynamic Interface Exploration and Learning. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (Pittsburgh, PA, USA) (UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 48, 17 pages. <https://doi.org/10.1145/3654777.3676356>
- [87] Andrew Walsh. 2012. Mobile Information Literacy: A Preliminary Outline of Information Behaviour in a Mobile Environment. *Journal of information literacy* 6, 2 (2012), 56–69. <https://doi.org/10.11645/6.2.1696>
- [88] Tianyu Wang, Giuseppe Cardone, Antonio Corradi, Lorenzo Torresani, and Andrew T. Campbell. 2012. WalkSafe: A Pedestrian Safety App for Mobile Phone Users Who Walk and Talk while Crossing Roads. In *Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications (HotMobile '12)*. Association for Computing Machinery, New York, NY, USA, Article 5, 6 pages. <https://doi.org/10.1145/2162081.2162089>
- [89] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, Kigali, Rwanda, 24 pages. <https://openreview.net/forum?id=1PL1NIMMrw>
- [90] Edward S Warner, Ann D Murray, and Vernon E Palmour. 1973. *Information Needs of Urban Residents*. US, Bureau of Libraries and Learning Resources, Baltimore, MD, USA.
- [91] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and Social Risks of Harm from Language Models. arXiv:2112.04359 [cs.CL] <https://arxiv.org/abs/2112.04359>
- [92] Hao Wen, Yuanchun Li, Guohong Liu, Shanhui Zhao, Tao Yu, Toby Jia-Jun Li, Shiqi Jiang, Yunhao Liu, Yaqin Zhang, and Yunxin Liu. 2024. AutoDroid: LLM-powered Task Automation in Android. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking (Washington D.C., DC, USA) (ACM MobiCom '24)*. Association for Computing Machinery, New York, NY, USA, 543–557. <https://doi.org/10.1145/3636534.3649379>
- [93] Apurwa Yadav, Aarshil Patel, and Manan Shah. 2021. A Comprehensive Review on Resolving Ambiguities in Natural Language Processing. *AI Open* 2 (2021), 85–92.
- [94] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review* 11, 12 (2024), nwae403.
- [95] Johanna Ylipulli, Tiina Suopajarvi, Timo Ojala, Vassilis Kostakos, and Hannu Kukka. 2014. Municipal WiFi and Interactive Displays: Appropriation of New Technologies in Public Urban Spaces. *Technological Forecasting and Social Change* 89 (2014), 145–160. <https://doi.org/10.1016/j.techfore.2013.08.037>
- [96] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling Context in Referring Expressions. In *European conference on computer vision*. Springer, Springer, Amsterdam, The Netherlands, 69–85.
- [97] Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. 2023. Conversational Information Seeking. *Foundations and Trends in Information Retrieval* 17, 3–4 (Aug. 2023), 244–456. <https://doi.org/10.1561/15000000081>
- [98] Nima Zargham, Mateusz Dubiel, Smit Desai, Thomas Mildner, and Hanz-Joachim Belz. 2024. Designing AI Personalities: Enhancing Human-Agent Interaction Through Thoughtful Persona Design. In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia (MUM '24)*. Association for Computing Machinery, New York, NY, USA, 490–494. <https://doi.org/10.1145/3701571.3701608>
- [99] Junge Zhang, Feihu Zhang, Shaochen Kuang, and Li Zhang. 2024. NeRF-LiDAR: Generating Realistic LiDAR Point Clouds with Neural Radiance Fields. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 7 (Mar. 2024), 7178–7186. <https://doi.org/10.1609/aaai.v38i7.28546>
- [100] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. Association for Computing Machinery, New York, NY, USA, 177–186. <https://doi.org/10.1145/3269206.3271776>
- [101] Haoyi Zhu, Honghui Yang, Yating Wang, Jiange Yang, Limin Wang, and Tong He. 2025. SPA: 3D Spatial-Awareness Enables Effective Embodied Representation. In *The Thirteenth International Conference on Learning Representations*. OpenReview.net, Singapore, 31 pages. <https://openreview.net/forum?id=6TLdqAZgzn>

## A Participants' Initial Goal of the Walking Session

**Table 4: Participants' initial goals of the walking session. NA indicates that the participant started with exploratory wandering rather than a specific purpose.**

Participant	Initial Goal
P1	Find an Italian restaurant for lunch
P2	Visit a new park
P3	NA
P4	Find a sunset spot
P5	Pick up Museum Night wristband
P6	NA
P7	NA
P8	Find a secondhand clothing store
P9	NA
P10	Visit a stadium
P11	Visit a graffiti wall
P12	Find a vegan lunch place
P13	NA
P14	Buy fresh blueberries in a market
P15	Visit the sky bar building
P16	Find a pink house mentioned by a friend
P17	NA
P18	NA
P19	NA
P20	Buy flowers for the balcony
P21	NA
P22	Find a nail salon
P23	NA

## B Walking Region for the Technology Probe Session



**Figure 8:** Map showing the approximate walking region for the technology probe session.

### C Accuracy of UrbanSearch Responses

To analyze the accuracy of UrbanSearch’s responses to the participants’ queries collected in the study outlined in Section 3, we recruited two residents of Tampere, who had lived in the city for at least two years, could communicate fluently in English, and held at least a graduate degree. We excluded 14 queries that served only greetings (e.g., “hello”) or emotional expressions (e.g., “cool!”) and anonymized 588 (user query, system response) pairs with the necessary context (e.g., photos, geolocation) before sharing them with the evaluators as Excel files. When determining the correctness of each query response, the evaluators could conduct online searches and map searches, as well as use their own knowledge about the local neighborhood. We instructed them to annotate each response with “correct,” “incorrect,” and “no answer given.” We explained that the label “no answer given” was used for responses such as “I cannot tell.” We also invited evaluators to provide free-form comments on the overall quality of the responses. We gave them one week to complete the annotation.

After both evaluators completed the annotations, we recruited a third evaluator with the same qualifications to perform annotations for (user query, system response) pairs for which the two evaluators disagreed. We merged the annotations by taking the majority vote; all conflicts were successfully resolved through this process.

Excluding the 32 responses for which UrbanSearch did not provide an answer (“no answer given”), UrbanSearch generated 527 (94.7%) “correct” responses and 29 (5.2%) “incorrect” responses. Notably, queries related to geospatial tasks led to a significant proportion of queries labeled either as “no answer given” (27 of 32) or “incorrect” (22 of 29). Evaluator 1 commented “It knows separated places in the area but does not have a whole picture or a dynamic spatial awareness.”

### D Building Initial Trust Through Known-Answer Queries and Ongoing Calibration

Interestingly, most participants ( $n = 19$ , 82.6%) began their interactions by posing questions they could verify—either from prior knowledge or through external tools—as a way to test the probe’s capability. We observed that participants’ prior exposure to AI tools strongly influenced how they framed their capability-testing queries. Regular or daily users tended to probe for functions they perceived as novel or advanced compared to their past experiences. For instance, P4, a daily ChatGPT user, captured an image of his surroundings and asked, “What’s this place about?”—explicitly testing the system’s locational awareness, which he described as “unfamiliar but very important in the urban space.” By contrast, participants with little or no prior AI experience focused on more basic expectations, such as verifying common-sense tasks or combining image and text inputs. For example, P15, who had no prior exposure to AI tools, took a photo of a red light and asked, “Can I cross the street?”

We found that this initial capability testing helped participants build trust in UrbanSearch, leading them to accept its responses without seeking external verification for later queries. However, most of them ( $n = 21$ , 91.3%) reported having no clear mental model of how it worked. After the initial capability testing, P3 began trusting UrbanSearch’s capability and said, “*It gave me correct answers before.*” However, during the interview, the participant later reflected: “*Now I realize I sometimes took its responses by faith. It seems very intelligent, but I did not understand how it worked.*”

As interactions continued, our participants’ perceptions of UrbanSearch’s trustworthiness fluctuated beyond the initial capability testing. Participants perceived UrbanSearch more trustworthy when it openly acknowledged its capability limitations. For example, P19 took a photo of a passageway with a construction barrier and asked, “Is it a dead end?” UrbanSearch replied: “It looks like a passageway, but I can’t confirm if it’s a dead end.” P19 appreciated this conversation and remarked: “*I would trust the probe more because it let me know it cannot know something, which is better than pretending to know all.*” In contrast, inconsistencies in responses and the absence of transparent information sources often undermined participants’ trust in UrbanSearch. When responses contradicted their expectations, participants often expressed doubt. For instance, when P14 asked about the nearest tram station and doubted the initial reply: “Are you sure there is no other stop nearer?” UrbanSearch changed its answer on the second turn. P14 interpreted this shift as an inconsistency and subsequently avoided distance-related questions. Participants also explicitly sought evidence to support responses, as when P10 asked, “What are your sources for this knowledge?” Similarly, P1 reflected: “*I could not fully trust lines of text. I know it might hallucinate, and I expect to see links and maps as a reference.*” Many participants ( $n = 16$ , 69.6%) expressed heightened doubt in the absence of such transparency and voiced a strong preference for external references, such as Wikipedia links or maps, to help them assess the credibility of the response.

### E Interaction Threads Overview of Technology Probing Session

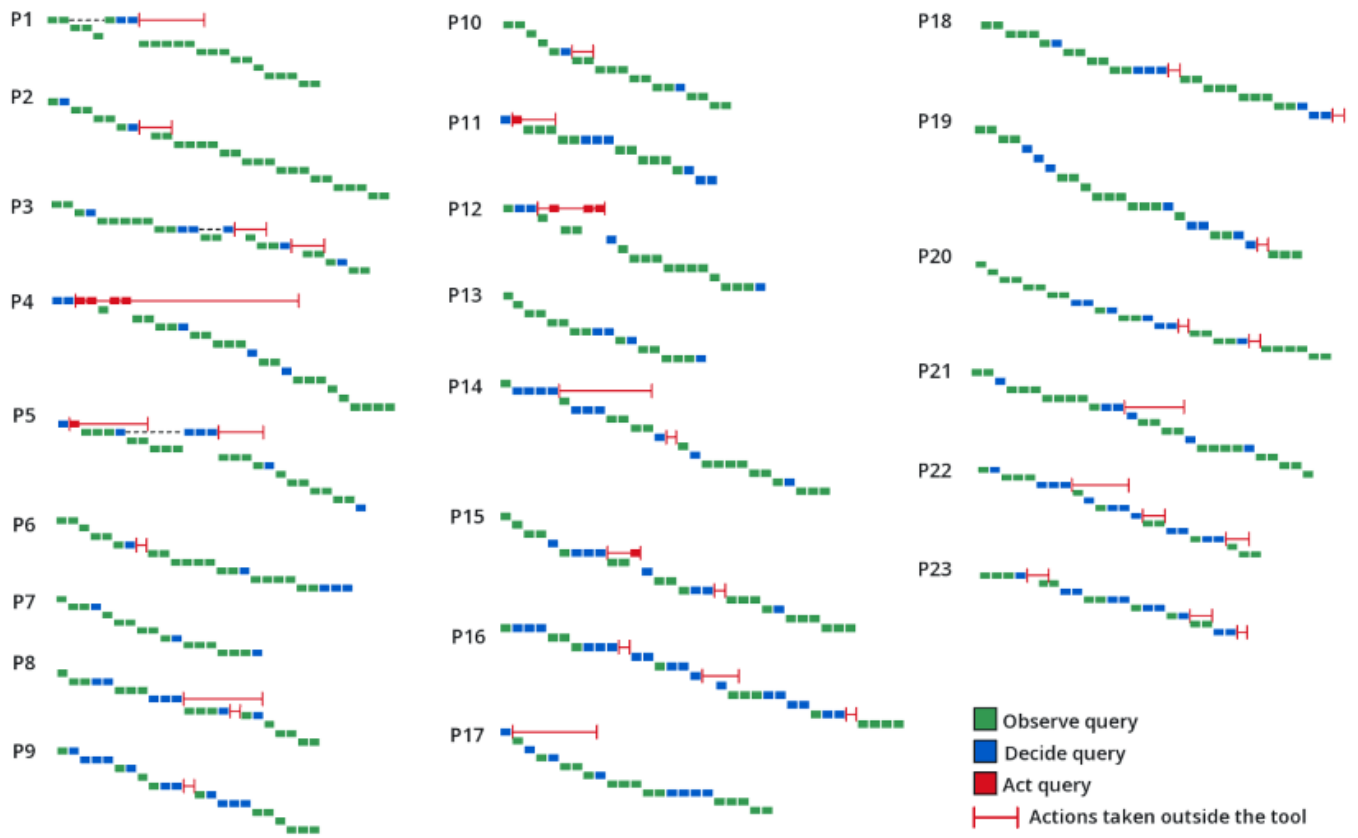


Figure 9: Participant-level interaction threads. Each row corresponds to one participant’s interaction thread, defined as a topic-based sequence of queries and actions. Squares represent queries and are color-coded as follows: green = observe, blue = decide, red = act. Red brackets indicate actions taken outside UrbanSearch (e.g., using Google Maps or making in-person purchases).