Probing the Gaps in ChatGPT's Live Video Chat for Real-World Assistance for People who are Blind or Visually Impaired

Ruei-Che Chang
rueiche@umich.edu
University of Michigan
Ann Arbor, MI, USA

Rosiana Natalie
rosianan@umich.edu
University of Michigan
Ann Arbor, MI, USA

Wenqian Xu
wxtu@umich.edu
University of Michigan
Ann Arbor, MI, USA

Jovan Zheng Feng Yap
jovanyap@umich.edu
University of Michigan
Ann Arbor, MI, USA

Anhong Guo
anhong@umich.edu
University of Michigan
Ann Arbor, MI, USA

Abstract

Recent advancements in large multimodal models have provided blind or visually impaired (BVI) individuals with new capabilities to interpret and engage with the real world through interactive systems that utilize live video feeds. However, the potential benefits and challenges of such capabilities to support diverse real-world assistive tasks remain unclear. In this paper, we present findings from an exploratory study with eight BVI participants. Participants used ChatGPT's Advanced Voice with Video, a state-of-the-art live video AI released in late 2024, in various real-world scenarios, from locating objects to recognizing visual landmarks, across unfamiliar indoor and outdoor environments. Our findings indicate that current live video AI effectively provides guidance and answers for static visual scenes but falls short in delivering essential live descriptions required in dynamic situations. Despite inaccuracies in spatial and distance information, participants leveraged the provided visual information to supplement their mobility strategies. Although the system was perceived as human-like due to high-quality voice interactions, assumptions about users' visual abilities, hallucinations, generic responses, and a tendency towards sycophancy led to confusion, distrust, and potential risks for BVI users. Based on the results, we discuss implications for assistive video AI agents, including incorporating additional sensing capabilities for real-world use, determining appropriate intervention timing beyond turn-taking interactions, and addressing ecological and safety concerns.

CCS Concepts

• **Human-centered computing → Empirical studies in accessibility**; **Accessibility technologies**.

Keywords

visual descriptions, blind, visually impaired, assistive technology, accessibility, live video, large multimodal models, real world

This work is licensed under a Creative Commons Attribution 4.0 International License.
ASSETS '25, Denver, CO, USA
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0676-9/2025/10
https://doi.org/10.1145/3663547.3746319

ACM Reference Format:
Ruei-Che Chang, Rosiana Natalie, Wenqian Xu, Jovan Zheng Feng Yap, and Anhong Guo. 2025. Probing the Gaps in ChatGPT's Live Video Chat for Real-World Assistance for People who are Blind or Visually Impaired. In *The 27th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '25), October 26–29, 2025, Denver, CO, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3663547.3746319

1 Introduction

Recent advancements in large multimodal models (LMMs) have significantly advanced assistive technologies, empowering blind or visually impaired (BVI) individuals to independently access, interpret, and interact with their environments. Traditionally, BVI users relied on sighted assistance by capturing images and requesting visual descriptions from online crowd workers [20, 86], or sought real-time guidance through remote sighted assistance (RSA) services like BeMyEyes [2] and Aira [1]. More recently, LMMs-powered applications, such as Be My AI [5] and SeeingAI [8], have enabled greater autonomy by allowing users to interactively obtain visual information directly from photos through AI-driven conversations. Prior research has explored how BVI individuals incorporate these emerging technologies into their daily routines, which offered valuable insights into both their practical utility and ongoing challenges for future development of assistive technologies [11, 16, 22, 31, 36, 72, 73, 75, 91, 92].

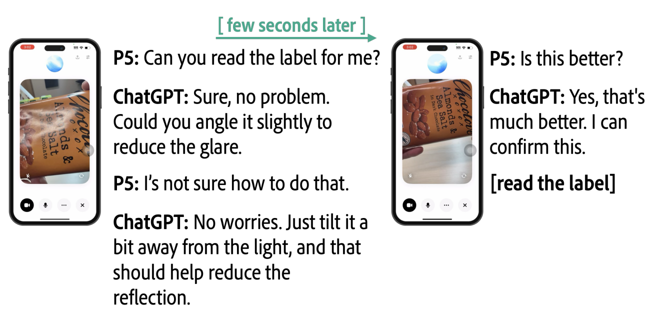


Figure 1: Illustration of user turn-taking interactions with ChatGPT's live video AI. The screenshots were cropped from the video recording of P5.

On December 12, 2024, OpenAI introduced ChatGPT Advanced Voice with Video[1] [3], a state-of-the-art live video AI system that enables visual question-answering (VQA) through live video feeds. This innovation allowed users to directly query and access the dynamic real world in real time, which eased the previous need and efforts to capture static photos and await responses. However, at the time of our study, this feature was still available only through a costly subscription-based service with daily usage limitations. Although a partnership between Be My Eyes [2] and OpenAI was announced, the limited availability of ChatGPT's live video AI for most potential users restricted opportunities to conduct contextual inquiries into its natural integration within their daily routines.

To address this limitation and investigate how live video AI could effectively assist BVI individuals in practical, everyday tasks, we conducted an in-person user study with eight BVI participants. Drawing from prior research, we created diverse task scenarios designed to engage participants with ChatGPT's live video AI. These scenarios included tasks such as identifying unknown objects, distinguishing multiple objects, recognizing prominent visual landmarks, and locating specific items or landmarks within unfamiliar indoor and outdoor environments. Specifically, we aimed to understand these questions:

RQ1: How do BVI users leverage ChatGPT to accomplish a variety of visual access tasks?
RQ2: How do BVI users perceive ChatGPT (and vice versa)?
RQ3: What are the limitations of ChatGPT that hinder its effectiveness?

We found that ChatGPT was able to provide hands-on, useful guidance and information for tasks, such as guiding users to aim at objects to reduce glare, reading labels, or identifying static visual information. Also, participants regarded ChatGPT as a tool to complement their own orientation and mobility (O&M) skills rather than a complete replacement for traditional assistive methods. However, participants faced difficulties due to ChatGPT's inability to provide live descriptions during dynamic activities, which forced participants into repetitive interactions.

Also, participants generally found interactions with ChatGPT realistic and human-like due to the fluid conversation nature. However, trust was undermined when ChatGPT incorrectly assumed visual abilities, which requested visual verification from its users or suggested seeking help from other sighted individuals. In response, users developed strategies to explicitly inform ChatGPT about their visual impairments, though ChatGPT consistently failed to fully adapt to these clarifications.

Finally, we identified several key limitations. For instance, ChatGPT's responses often relied on general knowledge rather than being grounded in users' immediate reality. Its tendency to respond overly affirmatively sometimes resulted in incorrect confirmations or potentially harmful guidance. Also, the lack of accurate visual memory and spatial awareness led to incorrect instructions and required users to generate redundant and repetitive prompts. Based on these findings, we discussed implications for designing a context-aware video AI agent towards real-world assistive use. This included enhancing mutual understanding and awareness between users and the AI's capabilities, integrating new sensing capabilities for real-world assistive perception, balancing general knowledge with specific information, and navigating issues of trust and realism in human-like AI assistance.

[1] We refer to this feature as ChatGPT throughout this paper.

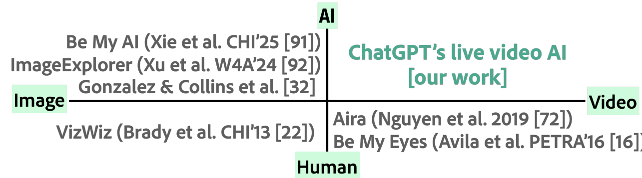


Figure 2: The scope of our work, situated in other literature that explored the assistive technologies used in the daily lives of BVI. We focused on how live video AI could be used for real-world assistive use.

2 Related Work

This work was inspired by the evolution of human and AI-powered VQA systems and the information needs of BVI users in different scenarios, which motivated our study task design.

2.1 From Human to AI-powered System for Real-World Accessibility

Traditionally, BVI individuals could capture a photo and request visual descriptions by seeking assistance from online crowd workers [20, 86]. Currently, RSA services, such as BeMyEyes [2] and Aira [1], connect BVI users with sighted agents through video calls, providing real-time verbal guidance to address dynamic requests. However, conveying visual information in real time can be challenging for sighted agents due to the high cognitive load and temporal demands involved in rapidly interpreting and verbally communicating visual environments [37, 43, 57, 58]. Additionally, RSA services can raise privacy concerns [23], incur high costs (*e.g.,* $65 per month for professional services such as Aira [1]), and volunteer-based options like Be My Eyes [2] may not always be available.

With recent advancements in AI, LMMs have revealed the potential to apply AI to the real-world applications previously unattainable without human intervention, such as navigation [48, 98], camera aiming [94], text input suggestion [63], image editing [26], and photo reminisce [93]. Commercial assistive applications, including Be My AI [5] and SeeingAI [8], enable users to obtain answers to queries about photos or images and BVI users could also explore the descriptions more interactively, such as using touch-based exploration [4, 50, 71]. Recently, AI-powered assistive technology has evolved beyond static images toward dynamic media. Applications such as SeeingAI [8] and PiccyBot [7] have introduced features for providing video description, and also allowed BVI users to ask follow-up questions [7, 83]. However, these solutions still face limitations in rapidly changing, real-time scenarios that necessitate immediate and dynamic description generation. WorldScribe [25], developed to provide live visual descriptions during real-world exploration, may not always sufficiently adapt to the evolving needs and intents of users navigating dynamic environments.

In December 12, 2024, OpenAI introduced the "ChatGPT Advanced Voice with Video" [3], enabling VQA through live video feeds. This allowed users to dynamically converse with AI according to their immediate informational needs and evolving intents. Unlike image-based non-voice systems that offer detailed descriptions, video-based interactions present short, immediate conversational responses with a human-like voice and tone. This opened both new possibilities and challenges for independently engaging with their real-world visual surrounding without the sighted assistance. Hence, different from prior studies that explored commercial AI-powered photo-based technologies for BVI users (e.g., Orcam [6, 32], BeMyAI [19, 91], and SeeingAI [32, 47]), our research investigates how ChatGPT's new capability can enhance assistive interactions beyond traditional photo-based VQA methods.

2.2 Information Needs of BVI People in Different Scenarios

BVI individuals face a number of challenges in everyday tasks. Early research analyzing over 4,000 photo-based questions [22] found that BVI users often struggle with tasks ranging from basic activities such as locating objects [12, 31, 67] and reading object labels [21, 22, 54], to more complex ones like identifying visual landmarks [24, 31] or interpreting scenes with intricate spatial relationships [100]. Moreover, the information needs of BVI users vary depending on both the context of access and the type of image content (e.g., object versus scene) [31, 45, 68, 84, 85]. For example, during grocery shopping, users may need to locate a product and read associated text labels (e.g., expiration dates, nutritional facts), but also rely on contextual visual cues not explicitly labeled, such as signs of damage or wear (e.g., scratches). Therefore, in this paper, we designed a range of object understanding scenarios, including object localization, comparison, and label reading, to better support the diverse needs of BVI individuals for real-world information access (Table 2).

In addition to object understanding, BVI users have critical information needs related to their surrounding environments to support safety, independence [18, 42], and efficient navigation [9, 33, 59]. These needs could vary significantly depending on whether the context is indoor or outdoor. Traditional assistive navigation technologies often struggled to operate seamlessly across both settings [33, 46, 70, 82, 90], particularly due to the limitations of GPS in indoor environments [28, 42, 78]. Indoor spaces tended to be more complex, characterized by multiple floors, narrower pathways, and denser obstacles, which necessitated specialized navigation strategies [69, 78]. Studies have shown that in indoor environments, BVI individuals built spatial awareness and memorized routes by associating landmarks [9, 14, 15, 88], and preferred egocentric directional cues such as heading, direction, and distance traveled [64, 88]. In contrast, when navigating outdoor environments, BVI users prioritized essential wayfinding details such as landmarks, routes, crossing points, and building features like entrances [22, 40], as well as real-time information regarding obstacles, current location, orientation, and accessibility features [9, 12, 59]. To date, many of these complex tasks still require assistance from sighted individuals, primarily due to their high-stakes implications for safety [37, 43, 57, 58]. Motivated by this challenge, we explore whether live video powered by LMMs can provide comparable support to that of human assistance, and what gaps still remain. To investigate this, we present a set of scenarios centered on understanding visual landmarks and navigating toward them, tasks that GPS alone cannot adequately support.

In summary, building on prior work, we designed four tasks focused on object understanding and five on navigation (Table 2). These scenarios varied across several dimensions, including visual complexity (e.g., the number of objects in the object understanding tasks), intent ambiguity (e.g., *specific* vs. *general*), location (e.g., indoor vs. outdoor), and spatial scale in an indoor environment (e.g., a room vs. a floor). For object understanding, the tasks involved understanding an object and describing it as *general* intent, as well as comparing fine-grained differences between similar items as *specific* intent. For navigation, we developed tasks centered on visual and spatial understanding as *general* intent and landmark localization and navigation as *specific* intent. We also considered differences in navigation between *indoor* and *outdoor* environments. Specifically, within indoor settings, we explored tasks at *low* (e.g., within a room) or *high* spatial complexity (e.g., across a floor).

3 Study Method

We conducted an in-person user study involving eight BVI participants to explore how ChatGPT's Advanced Voice with Video [3] could support a variety of their daily activities. We used this system as the primary tool, since, as of 2025, it represents the state-of-the-art in LMM-powered video interactions, offering the ability of visual understanding and high interactivity by voice. Specifically, we aimed to answer these research questions:

RQ1: How do BVI users leverage ChatGPT to accomplish a variety of visual access tasks?
RQ2: How do BVI users perceive ChatGPT (and vice versa)?
RQ3: What are the limitations of ChatGPT that hinder its effectiveness?

3.1 Participants

We recruited eight BVI participants (3 Male and 5 Female) using email lists for local accessibility organizations, prior contacts, and snowball sampling. Participants aged from 18 to 72 (Avg. 45.5) and described their visual impairment as blind (N=6) or having low vision (N=2). Most participants had prior experiences using remote sighted assistance (RSA) and AI-enabled services, such as Orcam [6], BeMyEyes [2], BeMyAI [5], Aira [1], or SeeingAI [8] in their daily lives (Table 1).

3.2 Procedure

After obtaining their informed consent, we asked participants about the use of assistive technologies in their daily life, including navigation apps, text or image recognition apps, and other visual access tools. They were then introduced to ChatGPT and instructed on how to mute and unmute their phones, allowing them to speak with researchers privately without the AI overhearing.

The study involved certain scenarios conducted in both indoor and outdoor environments (Table 2), requiring participants to walk and explore. In each scenario, participants used ChatGPT to request visual information necessary for completing the task. They were encouraged to share feedback at any time. Additionally, researchers asked follow-up questions after participants either completed a

Table 1: Participants in our user study were marked as P1-P8.

ID	Age	Gender	Self-Reported Visual Ability	Assistive App Use
P1	41	Male	Blind due to Retinitis Pigmentosa, left < 0.5 degree, depends on lighting to identify the color of the object.	SeeingAI, BeMyAI, BeMyEyes, Aira, Orcam, SoundScape, and VoiceVista
P2	58	Female	Right: blind. Left: Usable vision using a physical magnifier.	SeeingAI, BeMyAI, BeMyEyes, Aira, and Orcam,
P3	50	Female	Blind, since birth. Light perception.	SeeingAI, BeMyAI, BeMyEyes, Aira, Orcam and BlindSquare
P4	73	Female	Blind, since birth. Light perception.	SeeingAI, BeMyAI, BeMyEyes, and Aira
P5	41	Male	Blind, since birth. Light perception.	SeeingAI, BeMyAI, BeMyEyes, and SoundScape
P6	60	Female	Blind, since birth.	BeMyAI and BeMyEyes
P7	24	Female	Blind, acquired since 13.	None
P8	18	Male	Low vision due to Stargardt. Right: 20/1000, Left: 20/600, Light to Moderate color blindness.	SeeingAI

scenario or chose to discontinue it. These questions focused on identifying unclear information, perceived errors, the effort needed to obtain useful responses, and potential improvements or desired solutions.

During the study, participants were accompanied by two researchers while engaging in specifically designed real-world scenarios. Participants were allowed to request assistance anytime from the researchers if they were unable to acquire necessary details from ChatGPT. Given the mobile and exploratory nature of the study, researchers ensured participants' safety by intervening when necessary.

To investigate ChatGPT's potential in supporting BVI users across diverse real-world contexts, we designed everyday scenarios with varying levels of complexity. As mentioned in Section 2.2, these scenarios differed based on factors such as visual complexity (e.g., number of objects in the object understanding task), intent ambiguity (e.g., specific vs. general), location (e.g., indoor vs. outdoor), and spatial complexity in the indoor environment (e.g., a room vs. a floor). Note that our goal was not necessarily for participants to complete every task, but rather to provide meaningful opportunities for them to engage with ChatGPT purposefully. We detailed the task setup and instructions we gave to our participants in Table 2.

The study took two hours in total, and each participant was compensated with their transportation costs and $50 for their participation. This study was approved by the IRB in our institution.

3.3 Apparatus

ChatGPT's new feature, "Advanced Voice with Video," was released on December 12, 2024, as a subscription-based service. However, its high cost and daily usage limits reduced its accessibility. Participants were briefly introduced to the feature and engaged in a short conversation with it before proceeding to the main scenarios. Participants held the phone with a clip-on guide to assist in locating the mute button during the study, allowing them to quickly mute ChatGPT while speaking with us and unmute it when they wished to interact with ChatGPT.

3.4 Data and Analysis

All interaction sessions between users and ChatGPT were logged. And the contextual inquiries were also video recorded and transcribed for analysis. The first author independently performed open coding on all transcripts to identify initial codes. The initial coding focused on the participants' strategies of using ChatGPT, participants' feedback on ChatGPT's responses, and any friction they experienced. Another two authors reviewed all coded excerpts and iterated on the set of codes through discussion with the first author. They then conducted affinity diagramming [35] on the initial codes to extract and organize high-level themes.

4 Results

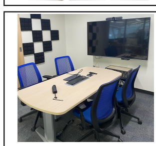
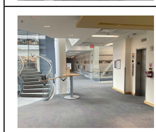
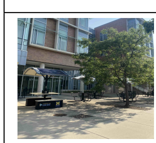
In this section, we present our study results organized around the research questions. First, we explore ChatGPT's capabilities and examine how participants utilized it across various scenarios. Next, we focus on how ChatGPT guided users through tasks by providing different types of instructions and information. Finally, we discuss participants' overall perceptions of ChatGPT and investigate its assumptions about its intended users.

4.1 How do BVI users leverage ChatGPT to accomplish a variety of visual access tasks?

In this section, we described how participants utilized ChatGPT across various scenarios. Specifically, we provided our observations of ChatGPT's effectiveness in providing suggestions for guiding visual search tasks, discussed participants' needs and strategies for prompting it to deliver live descriptions, and described how it could be useful to complement their O&M skills.

4.1.1 Effective suggestions for guiding visual searches. We observed mixed feedback from ChatGPT, which influenced task outcomes both positively and negatively. First, ChatGPT successfully answered specific questions when the visual scene provided was static and contained the relevant information the user sought. For instance, P8, who could not see the object details, showed chocolate bars with visible cocoa percentage labels and asked, *"Which of these 3 has the most cocoa in it."* ChatGPT responded accurately: "Based on the labels. The bar on the left has 70% cocoa. The middle one has 55%,

Table 2: Setup and instructions for each scenario. These scenarios differed based on factors such as Visual Complexity in object understanding tasks marked as *Low*, *Middle*, and *High* in red, Intent Ambiguity in blue (e.g., *Specific* vs. *General*), *Location* in orange (e.g., *Indoor* vs. *Outdoor*) and Spatial Complexity in indoor environment in purple (e.g., *Low* vs. *High*).

Image	Scenario	Setup	Instruction to User	Dimensions
	Understanding an object	Participants were given a cup with colorful graphics and texts.	You got a gift from your friend who just traveled back from a tourist spot. Can you use ChatGPT to understand this object? In terms of color, texts, and graphics.	**General** **Low**
	Understanding and distinguishing two different spice bottles	Participants were given two spice bottles from Trader Joe's, including one chili lime seasoning with a red label and lid, and another oregano with a green label and lid.	In the grocery store, you have two spice bottles with different labels, colors, and texts. Can you use ChatGPT to tell the differences and the similarities between them?	**General** **Low**
	Understanding and categorizing four spray bottles	Participants were given two identical (from the brand *Everyone*, ruby grapefruit), and the other two were from the same brand (*Whole Foods 365*) but had different scents (cucumber aloe and lavender)	You just got the four spray bottles from a shared storage in your home. Can you use ChatGPT to categorize them based on their brands and scents?	**Specific** **middle**
	Finding products with specific information	Participants were given three carton of juices, including two apple juices (100 & 35 calories) and one lemonade (100 calories), and three chocolate bars (55, 65, 70% of cocoa).	You want to find some snacks in a shared pantry, specifically, the chocolate bars with the most cocoa and the apple juice with the fewest calories for your health. Can you use ChatGPT to help you find them?	**Specific** **High**
	Locating an umbrella in a room	A folded umbrella was put on the table in front of participants, who were told the umbrella was in the room.	Your partner is at the entrance of the first floor and waiting for you to take the umbrella for her. Can you use ChatGPT to help you find the umbrella in the room?	**Specific** **Indoor** **Low**
	Locating stairs or elevators on the floor	Participants were positioned in the middle of the hallway, where the elevator and stairs were visibly on their left, 20 feet away.	Now you have the umbrella, can you use ChatGPT to help locate the stairs or elevator on this floor?	**Specific** **Indoor** **High**
	Understanding the indoor environment	Participants started in the atrium on the first floor of our building, with all floors, spiral stairs, and the skyline visible.	Now you are on the first floor. You want to share your experiences here after the study with your sighted friends. Can you use ChatGPT to understand the building's interior design, which could help spark a conversation with them?	**General** **Indoor**
	Locating a sheltered outdoor reading spot	participants started from the outside of one building while sheltered benches were visibly located at the entrance of the next building.	You are new here and want to find a peaceful outdoor spot with a bench and a cover to avoid sunshine or rain to enjoy reading. Can you use ChatGPT to help locate and navigate to it?	**Specific** **Outdoor**
	Understanding surroundings for rideshare driver pickup	Participants started from the sheltered bench, surrounded by buildings, swings, sidewalks, and a parking lot is far away but visible.	You just called the Uber driver and want to know where you are so that you can describe to your driver how to help find you. Can you use ChatGPT to help you understand the current environment to support this?	**General** **Outdoor**

and the one on the right has 65%. So the left bar has the highest cocoa content." Instead of simply answering direct questions, ChatGPT also offered useful guidance for problem-solving. For example, P5 attempted to read the information on the chocolate bar (Figure 1):

> **P5**: *Can you read the label for me?*
> **ChatGPT**: *Sure, no problem. Could you angle it slightly to reduce the glare?*
> **P5**: *I'm not sure how to do that.*
> **ChatGPT**: *No worries. Just tilt it a bit away from the light, and that should help reduce the reflection.*
> **P5**: [After tilting the chocolate] *Is this better?*
> **ChatGPT**: Yes, that's much better. I can confirm this. [read the information on the label]

Additionally, while participants searched for an umbrella in the room, ChatGPT provided various interactive and helpful suggestions. For instance, ChatGPT suggested to P4, "Let's start by checking near the door, or maybe under the table. Could it be there? ... Maybe in a bag or drawer." P4 appreciated these ideas and noted that *"it gave good suggestions that they could have fallen down or in a bag."*

4.1.2 Steering ChatGPT to describe lively. ChatGPT enabled real-time VQA based on a live video feed. However, we observed that it provided information only when explicitly prompted by the user and did not proactively deliver descriptions, even when participants requested continuous feedback. All participants attempted to guide ChatGPT toward generating live visual descriptions across various scenarios but were unsuccessful. For instance, one participant tried to obtain information about a spice bottle by asking ChatGPT to describe it continuously while rotating the object:

> **P1**: *Okay, I'm going to rotate the one on the left. Tell me when you can read the main label.*
> **ChatGPT**: *sure go ahead and rotate it. I'll let you know when I can see the main label. [then kept silent]*
> **P1**: *Well, I spun it all the way around, and you never stopped me. So I'm guessing you can't really read the main label.*

Other participants also tried similar strategies in various other scenarios, such as *"Can you tell me when you see the umbrella?"* while searching for an umbrella - P5, or *"Can you tell me when there are obstacles in the way? ... Tell me when you see signs, or when you see the elevator doors."* while navigating an environment - P4. Most participants (P1, P3, P4, P5, P6, P7) also questioned ChatGPT's capability to provide live descriptions, explicitly asking it to confirm its ability to describe the environment lively and receiving deceptive affirmations:

> **P5**: *I don't think you're taking pictures on a continuous basis. So as I move the camera, you won't be describing anything new to me unless I ask you a specific question. Correct?*
> **ChatGPT**: *I can see the camera feed in real time. So I can describe what I see as you move. No need to ask specific questions. Just let me know what you need help with.*
> **P5**: *Go ahead and just describe what you see as I move the camera*
> **ChatGPT**: *All right. Now I see the black acoustic panels on the wall, a large screen and part of the table with a coffee cup. If there's a specific area you want to explore, let me know.*
> **P5**: [After panning the camera around and ChatGPT did not respond at all] *See? It's lying to me.*

Due to this limitation, participants adapted by frequently repeating short questions. For example, P3 prompted ChatGPT eighteen times with, *"Do you see the umbrella?"*, until finally locating it. Participants emphasized the need for live descriptions in mobile scenarios rather than repeated prompting, as P6 explained: *"I had to keep asking questions to get additional information. Because I'm in an unfamiliar area, I needed to reassure myself that I was still headed the right way. Make sure it's following around things that are obviously like walkways or whatever. So verifying that."*

Also, the lack of efficient memory on previous contexts could also make tasks that required repetitive actions more time-consuming as it did not fully comprehend prior user prompts to provide continuously support (Section 4.1.2), as described by P1, who tried to understand each spray bottle with similar prompts repetitively: *"You have to continually read them [spray bottles], and see what they say. Maybe having it remember what I'd like to ask when I do it again and again is more convenient."*

4.1.3 Needs and ways to leverage visual descriptions to complement O&M skills in different contexts. During our study, we observed that directional information provided by ChatGPT was mostly inaccurate or unclear, a concern also raised by all participants. For instance, ChatGPT stated, "I see a staircase behind you that could be a good option to get to the 1st floor," even though P2 was aiming the camera forward and there was no staircase behind the user based on our video analysis. Also, P7 remarked on the limited usefulness of ChatGPT's directions after trials and errors: *"I was not following it. I didn't go against its directions, but I just don't think it gave me really specific directions."* We also found that when entering video mode on a smartphone, ChatGPT displays an explicit warning on the screen: "Don't use live navigation or make decisions that may impact your health or safety.", which flashed out quickly and was inaccessible to screen reader users. Given this open limitation, in this section, we primarily discuss the potential role of ChatGPT as a visual prosthetic aid to complement personal navigation skills, rather than as a standalone navigation tool comparable to outdoor GPS-based or indoor beacon-based systems.

Due to constraints imposed by the turn-taking interaction approach, participants frequently prompted ChatGPT to describe their visual surroundings during navigation tasks, including identifying obstacles or signs. For instance, P6 prompted ChatGPT when searching for the umbrella in the room: *"Is there any barrier between me and the wall?"*, with her reason behind this query: *"There's always all kinds of detritus left around in hallways, or people drop things. I could slip and fall, so I want to know if it's a clear path to go."* P1 further suggested enabling proactive identification of relevant signs: *"If it knows I'm looking for stairs and sees a sign for stairs, it should say, I see a sign for stairs up ahead,' or maybe I see a sign for an elevator."'* Participants also expressed that visual descriptions could complement existing GPS systems. As remarked by P3, *"A lot of apps either do one thing and don't do the other. GPS systems will just* *navigate you and tell you how to get to a place. But they're not really describing as it goes."*

Despite the limitation in providing accurate directions, visual descriptions provided by ChatGPT could complement participants' existing skills or sensory information. For example, P5 utilized information from ChatGPT to supplement, but not replace, his mobility skills: *"I used my own auditory cues to supplement what ChatGPT was saying, because it might be very dangerous for a blind person to think this is my mobility need. And this is how I'm going to get around. We always want to understand that these are aids to mobility. But they're never going to replace good old-fashioned human mobility skills."* P2, who had limited usable vision, also benefited from ChatGPT's visual descriptions when finding the entrance of a building, with guidance involving lighting:

> **P2**: *Can you help me find the front door?*
> **ChatGPT**: *Of course. The main entrance should be straight ahead past the seating area, look like the glass doors.*
> **P2**: *I can't necessarily see the glass doors. Is there any other direction you can give me? Or is there a landmark I may be able to pick up?*
> **ChatGPT**: *I see a lot of light, head towards that bright area. The doors are right there.*

Afterwards, P2 successfully reached the entrance by the pointer of light and her auditory sense: *"This is useful. I'm also listening to see if I hear someone come in. You know, the other sense kicks in."* Similarly, P6 noted that visual descriptions complemented auditory cues: *"If I need it more for the visual than the audio. Like here, there are people talking who I obviously can hear. But if you go there, you want to avoid hitting or running into the seating. And that's what I need to know."*

In terms of *location* and *spatial complexity*, we observed that during indoor tasks, participants adopted distinct navigation strategies using mobility aids and tailored their ChatGPT prompts accordingly. In rooms, they often used their hands to trace surfaces (e.g., walls, tables) and frequently asked questions such as *"Is the umbrella here?"*, along with occasional checks for obstacles like *"Is anything obstructing my path?"* In hallways, by contrast, participants, aided by a white cane, seldom asked about obstacles that could be detected by the cane. Instead, they focused on prompting ChatGPT to identify elevators or relevant signs. Furthermore, indoor navigation involved more frequent micro-level prompts related to nearby objects and spatial layout, whereas outdoor navigation emphasized broader spatial orientation and wayfinding.

4.2 How do BVI users perceive ChatGPT (and vice versa)?

In this section, we explored participants' perceptions of ChatGPT, as well as how ChatGPT perceived participants. Specifically, we discussed the reasons behind participants' tendency to treat and perceive ChatGPT as human, the strategies to coach ChatGPT to recognize their abilities due to its frequent assumption of users' visual abilities, and ChatGPT's capacity to provide descriptions using non-visual language.

4.2.1 Users treated ChatGPT as human. In general, participants perceived interactions with ChatGPT just like interacting with a human due to several advantages, such as *"less efforts to take photos"* (P1), *"no delay or glitches compared to remote sighted assistance"* (P2), or *"a cheery and realistic voice and being able to talk back and forth"* (P5). ChatGPT also facilitated the conversation by frequently asking questions back, such as "I see a room with a chair and some black soundproofing panels in the wall. Are you in a studio?" after P3 asked for locating the umbrella, or "[info about the chocolate bar], what do you prefer?" after P4 asked for the cocoa content of the chocolate bars. However, as mentioned in previous sections, participants also lost trust in ChatGPT when perceiving its mixed, confusing responses along the way, which were not aligned with their expectation of how a human-like system should be, such as *"It's kind of tricks you into thinking it's a human being. And so you get that false sense of security"* (P5), *"if the AI were more specific, it would be helpful, but, as is now, I wouldn't use them"* (P4) or *"Don't really trust it. Cause I will put my safety above convenience every time"* (P6).

4.2.2 User strategies to address visual assumptions in ChatGPT. We identified several instances where ChatGPT inappropriately asked participants to rely on their own vision to verify visual content. For example, when ChatGPT was unable to access information from a bottle's label, it directed users to read it themselves, asking questions like "Are you able to read them off the label?" when P4 requested label details, "Are there any signs or doors in front of you?" when P2 sought an entrance, "Do you notice any signs or landmarks?" when P6 requested information about her surroundings, "you can check the label for the exact weight." when P3 accessed the label of spice bottles, or "Do you see any closet or storage area?" when P7 searched for an umbrella in the room. Such prompts caused confusion and frustration for multiple participants (P1, P2, P3, P4, P6, P7) as P7 remarked: *"This was trained for sighted people but not blind people."* Moreover, ChatGPT occasionally suggested that participants seek assistance from sighted individuals. For instance, it advised P2: "You might have to check a sign or ask someone nearby." Reflecting on this interaction, P2 commented: *"Do you think someone who is blind or visually impaired might feel dismissed when the AI tells them to ask someone else? It might feel like the system isn't truly adapted to their disability. Nobody really wants you to know that they cannot see because they are always afraid that people will say this to me."*

Noting that ChatGPT was not originally trained with visually impaired users in mind, several participants developed their own strategies to help ChatGPT better understand their disability and offer more suitable guidance. For example, P1 leveraged his previous experience coaching a photo question-answering system, where he needed to explicitly mention his visual impairment to receive descriptions involving people; otherwise, LMMs typically avoided describing humans due to privacy concerns. Consequently, he began specifying prompts at the beginning of each conversation, such as *"GPT, I am blind and I am in an unfamiliar room trying to find an umbrella."* Despite this, ChatGPT did not offer additional useful descriptions, leading to P1's frustration, expressed as *"I want it to tell me what to do to help it help me."* This coaching approach was also adopted by other participants. For example, P7: *"I am blind. I need your help."* P2: *"I can't see the stairs. Is there any way you* *can help me with that?"*, or P4: *"I can't see, you tell me."* Specifically, P2, who had usable vision in her left eye but no sight on her right, hoped ChatGPT can complement her vision *"I'm visually impaired, and I don't have any sight on my right, can this be able to pick up and say there's a door in front of you on your right?"* P6 also mentioned having cane with her occasionally in her prompt, such as *"Do I follow the walk along the grass here? I got my cane."*

Among participants, P5 explored various methods to prompt ChatGPT about their disability, such as asking for non-visual descriptions:

> **P5**: *I'm trying to describe this cup using non-visual language for a blind person who has never seen.*
> **ChatGPT**: *That's a great idea. Let's start with the shape and size. Is it more like a typical coffee mug or something smaller?*
> **P5**: *Well, this is the cup I'm holding in my hands.*
> **ChatGPT**: *Got it. So it's a small cup, maybe an espresso size. We can describe its texture, color, and any unique details. What's the color or pattern on it?*
> **P5**: *You think I want to create a cup. I'm talking about the image that you're receiving through a camera.*
> **ChatGPT**: *Got it. I can see the cup has a colorful pattern. It looks like it has illustrations or landmarks on it. Can you feel any texture differences, or is it smooth?*
> **P5**: *I can feel texture differences*
> **ChatGPT**: *Great. That means the design might be raised or embossed. What else stands out to you? Maybe the handle, shape, or weight?*

Besides talking about textures, ChatGPT also occasionally responded with confusing descriptions of it perceiving the environmental sounds, while P5 elaborated his blindness and detailed instructions to ChatGPT:

> **P5**: *I'm blind, so you'll have to help me understand how to orient. For instance, now I'm facing the building. I'm aware of that, but I won't be aware of all the features and things around me. What are the things I can mention to my driver?*
> **ChatGPT**: *I can hear the sound. You're near a building with an interesting canopy, or another thing you can mention is any pillars or columns around you [ChatGPT assumed user has visual ability]. Buildings with pillars often stand up to drivers.*
> **P5**: *I'm curious, ChatGPT, when you say you can hear the environment. Are you actually processing sound? Or are you describing using auditory or echolocation?*
> **ChatGPT**: *I can't actually process sound. When I mentioned cells or features. I try to give helpful suggestions from that.*

P5 explained his reasons behind his prompts *"I like telling it the nature of my problem before I go into it, so that it has an idea and can automate a lot of this stuff."* Although these adaptations occurred occasionally, most of the time, ChatGPT still failed to deliver helpful guidance, frequently proposing visually oriented instructions suitable only for sighted users. As P6 pointed out *"I hope it learns what different disabilities are because I told it I can't see. It doesn't seem to comprehend."* Similarly, P1 remarked *"It needs to be coached on what to do, like as I walk down the hallway, please read any signs and tell me if you see a staircase. Because they didn't really know what to tell me."*

4.3 What are the limitations of ChatGPT that hinder its effectiveness?

In this section, we discussed additional major limitations we observed during the study, including ChatGPT's frequent reliance on general knowledge rather than specific visual details, its overly affirmative responses leading to incorrect confirmations or potentially harmful guidance, and its lack of spatial memory causing to repetitive queries and confusion during navigation tasks.

4.3.1 Responding with general knowledge or information not grounded in reality. ChatGPT could provide general world knowledge useful for scenarios with a general intent. For instance, ChatGPT provided background information on the text and graphic labels of the cup to facilitate P4's understanding:

> **P4**: *Tell me about the Cup.*
> **ChatGPT**: *It's a cute souvenir cup. I see it, says [City name], and has some fun illustrations. Do you collect them?*
> **P4**: *Sure. What are the illustrations?*
> **ChatGPT**: *It looks like there's the [landmark name] and some other iconic [City name] landmarks, plus some cute details. It's got a nice bright blue interior, too.*
> **P4**: [After flipping to another side] *What's on this side.*
> **ChatGPT**: *looks like there's a baseball and a [animal name] definitely capturing some [City name] vibes.*
> **P4**: *Alright, what do you think the baseball represents?*
> **ChatGPT**: *the [baseball team name]? The city's Major League baseball team.*

Other examples included when P3 asked for the spray bottle *"What size is the bottle?"* with ChatGPT responded: "2 ounce bottle, the standard travel size." The action of adding additional explanation was also found in previous work when users interacted with LMMs [91].

However, though providing general world knowledge could be beneficial to uncover unknown facts, this could also confuse users. For example, in the scenario of looking for an umbrella, it inaccurately advised P3 with "It's inside your bed." and similarly misled P1 with "Could it be on a coat rack?", despite there being neither a bed nor a coat rack present. Also, the general knowledge could hinder the efficiency of accessing specific information. For instance, P4 presented two seasoning bottles, each with four sides but no visible nutritional information on the side P4 presented, and wanted to understand the ingredients:

> **P4**: *Tell me about the ingredients.*
> **ChatGPT**: *Oregano is just dried. Oregano leaves, the chili lime blend usually has chili pepper, salt, citric acid for the tang, and sometimes sugar or garlic. It's a mix of spicy and citrusy flavors.*
> **P4**: *Can you tell me the ingredients that are on the bottles?*
> **ChatGPT**: *Sure, for the oregano. It should just list organic oregano for the Chili line. It typically includes chili, pepper, sea salt, citric acid, and sometimes sugar*

> *and lime oil. You can check the label for exact details.*
> **P4**: *Can you check the label for me?*
> **ChatGPT**: *Sure. Let's see about the oregano. It should say just organic oregano. The chili lime usually has chili, pepper, sea salt, citric acid, and sometimes sugar or lime oil. Are you able to read them off the label?*

In this conversation, instead of guiding users to the side of nutritional facts, ChatGPT provided the general knowledge, which did not address the user's immediate needs. And P4 further commented *"when he [ChatGPT] was asked for specifics, he gave generalities."* Similar examples included when P3 asked about the calorie content of juice cartons, ChatGPT replied, "Usually vary by brand and ingredients. It's typically around 150 to 170 calories per serving. Check the label for specifics." Similarly, ChatGPT guided P7 to the elevator when she was facing a wooden door, by suggesting, "The elevator should be behind the door," and upon further inquiry from P7, explained its reasoning as, "I assumed it based on typical building layouts, but if there's a sign or button panel, that would confirm it." Again, in this case, ChatGPT asked P7 to visually confirm.

4.3.2 ChatGPT's tendency towards sycophancy. Aside from incorrect guidance or general knowledge provided out of context, we also found that ChatGPT had the tendency to cater to participants [61], which could lead to potential harm. For instance, in the following conversation, P7 held the two spray bottles with both scents of ruby grapefruit. ChatGPT was guided by the user to the incorrect answer, even though it answered correctly at first:

> **ChatGPT**: *The two spray bottles are ruby grapefruit.*
> **P7**: *Are you sure it's not peppermint citrus?*
> **ChatGPT**: *Oh, if the label says peppermint citrus, then it's definitely that.*
> **P7**: *What does the label say?*
> **ChatGPT**: *The label says it's 99.9% effective against germs. The scent is definitely peppermint citrus, and it's a 2 ounce spray bottle.*
> **P7**: *Are you sure that's what the scent is?*
> **ChatGPT**: *Yes, the label says peppermint citrus. The scent names are usually printed on the front.*
> **P7**: *Then where did you get ruby grapefruit from?*
> **ChatGPT**: *I misspoke. My mistake. It's definitely peppermint citrus. Thanks for catching that.*

P7 reflected on her overall experiences in such situations: *"ChatGPT has a tendency to say yes to what you say. So if I'm like, Oh, are there signs? It assumes that there are signs, and it kind of tries to make something. I don't want to say make it up, but like try to shift the reality so that there are signs. Same thing with like, are you sure it is not peppermint citrus?"* ChatGPT also often gave encouraging and cheery responses, such as "You're almost there! Just a few steps forward" or "You're doing great! Keep going", while participants had been misled or given wrong guidance from it during the tasks. This positive tone could lead to potential harms to the user, as P3 remarked *"The feedback is always positive. I mean, positive is good. But in the instance where a person might be in a dangerous situation, if there's obstacles, it needs to be able to tell them that and not say, Oh, you're almost there. Just keep going straight ahead. And they're getting ready to step off."* Similarly, P5 pointed out such tone could be harmful *"It's kind of tricks you into thinking it's a human being. And so you get that false sense of security, of everything it says is right, but I was dead wrong on certain things."* Hence, participants indicated their preference to have ChatGPT respond with *"just yes or no, the most accurate portrayal of the reality"* (P7) or *"rather to have no information at all than misleading information"* (P5).

4.3.3 Lack of precise spatial memory important to certain tasks. In several instances, we observed that ChatGPT appeared to lack precise visual memory, such as spatial information, within each conversation session, as it could not recall places participants had already explored, leading to incorrect guidance or repetitive suggestions. P3 and P6 both encountered this issue when searching for the umbrella, as P6 stated *"Oh, it's trying to redirect me. It didn't make sense to me to double back where I've just been, and then it didn't even recognize it. I never found it with memory. Again. ChatGPT doesn't always know everything."* When walking and exploring the outdoor scene, P3 and P5 also observed that ChatGPT did not remember the routes they have been to and gave false suggestions as P3 said *"It should be able to understand that you've already turned around already facing the right way. And it said turn around and go back because it couldn't tell that we had already turned."*

5 Discussion and Future Work

Our results revealed that ChatGPT was perceived as realistic and human-like, capable of effectively addressing specific queries and providing useful guidance. However, we also observed several key limitations of the state-of-the-art AI live video capability, such as the constrained turn-taking nature of interactions, its inconsistent capability to provide accurate information, and its overly-positive responses, which BVI users sometimes perceived as harmful or misleading. Currently, BVI users adapt to imperfect AI outputs by seeking alternative AI resources [36] or assistance from sighted individuals [11]. However, in the long term, following the principles of ability-based design [89], it is essential to design and develop AI systems that have context-awareness for providing accurate information, transparently communicate their capabilities, foster appropriate levels of trust, and proactively accommodate user needs, rather than placing the burden on users to bridge the gaps. In this section, we discuss our lessons learned from the study and implications for designing a context-aware video AI agent towards real-world assistive use.

5.1 Current Capabilities and Limitations of ChatGPT Live Video

Our findings highlighted both strengths and weaknesses of current ChatGPT capabilities across various scenarios. For object understanding tasks, particularly when users explored the objects generally, ChatGPT effectively leveraged its general world knowledge to help users uncover new insights and facilitate object comprehension. Unlike traditional image-based systems that tend to generate detailed and long-form image descriptions, ChatGPT produced short and immediate responses that resemble natural human conversation. This brevity influenced participants' interaction style, prompting them to ask frequent follow-up questions while limiting the amount of information conveyed in each turn. This short-form interaction style enhanced fluidity by allowing users to adjust camera aiming and corresponding questions in real-time, thereby eliminating the need to retake photos (Figure 1).

However, in cases where the object was not distinctly presented within the view, ChatGPT struggled to consistently provide accurate answers or specific guidance, often resorting to general world knowledge instead (Section 5.3). And ChatGPT exhibited a turn-taking interaction style that only responded to queries, even when participants asked it to provide essential information proactively (Section 4.1.2). Regarding navigation tasks, our findings indicated that the inherent turn-taking nature of ChatGPT and its inability to proactively generate essential live descriptions hindered effective navigation. Specifically, ChatGPT struggled with dynamically detecting and describing barriers or signs to users. Moreover, its inaccuracies in providing directional guidance and distance estimations, combined with limited spatial memory, severely limit its current usability for navigation purposes.

Overall, ChatGPT's turn-taking interactions through live video demonstrated effectiveness in scenarios involving clear, static visual presentations, but fell short in dynamically changing or ambiguous visual contexts. Furthermore, its assumption of users' visual ability impeded the delivery of necessary information to BVI users (Section 4.2.2). Future research should draw insights from effective human assistance to improve AI assistive systems in areas such as system assumptions about user abilities, continuous real-time visual processing for live descriptions, comprehensive spatial understanding, interaction memory, and managing the trade-off between generality and specificity of information based on context. We discussed each point in detail below.

5.2 Ability Awareness and Mutual Understanding in Human-AI Systems

To date, ChatGPT's video AI model's capabilities remain unclear. The black-box nature of AI poses challenges to user trust and adoption [29, 60, 77, 79]. Clearly communicating *"what the system can do"* is fundamental to effective human-AI interactions [13], especially in assistive technologies, where users with diverse abilities depend on predictable interactions for decision-making [11, 91]. This was also found important in the literature on conversational agents [65], which highlights how unmet or unclear expectations about system capabilities can lead to user frustration and disengagement. Our study results revealed several instances of such friction that could be mitigated through clear disclosure of abilities from both sides. For instance, participants frequently questioned the core technical capabilities of ChatGPT by asking specific technical questions, such as inquiring about its ability to provide live descriptions (Section 4.1.2), describe in non-visual language, and perceive sounds (Section 4.2.2). These were considered fundamental in human assistance [58]. When interacting with humans, such assessments of ability are often made through conversational cues and context, but the absence of analogous signals in AI systems makes it difficult for users to form accurate mental models of the system's capabilities [65]. Explicitly communicating AI capabilities or exposing its internal reasoning [62] during conversation may help users better understand system limitations and reduce misaligned expectations.

Conversely, from ChatGPT's perspective, we observed multiple instances where it mistakenly assumed BVI users had visual abilities, prompting them to verify visual content independently (Section 4.2.2). Consequently, BVI users were repeatedly compelled to explicitly clarify their visual limitations in each conversation, in order to *coach* ChatGPT to adapt to their needs (Section 4.2.2). Profiling and understanding BVI users' visual abilities and information preferences are already established practices within professional RSA services [43, 57]. Drawing from ability-based design principles [89] and these existing human-centered practices, future AI-powered assistive systems should develop the capability to be prompted with users' abilities, and then retain and provide adaptive information correspondingly, to facilitate more personalized and effective human-AI interactions in real-world assistive use.

5.3 General World Knowledge vs. Specificity

In our study, we found that ChatGPT was capable of providing helpful guidance and relevant information at both general and specific levels of detail (Section 4.1.1). However, there were instances where the timing or relevance of its responses was misaligned (Section 4.3.1). For example, under circumstances of limited or unclear visual input, it often defaulted to general world knowledge or made assumptions not grounded in the user's immediate visual context, specifically, when the label on a bottle was improperly presented (P3, P4), or when only a door was visible while the user searched for an elevator (P7). This problem could stem from the bias from training data, where BVI users tended to ask more urgent and less subjective questions than sighted people tended to ask [22, 61]. Despite these observations, we did not obtain sufficient evidence or consistent patterns regarding contextual factors influencing whether ChatGPT responded with general or specific information. This unpredictability can be problematic in live contexts, where users often rely on timely and precise feedback to complete real-world tasks efficiently and safely [22], which is still mostly tackled by RSA services [37, 43, 57, 58]. Ideally, inspired by human assistance, an AI assistive system should dynamically adjust its level of detail according to situational cues, prioritizing precise, contextually grounded information when the task demands it, and resorting to general knowledge only when appropriate or requested by users. Future assistive AI systems could achieve the balance between specificity and generality by improving visual scene understanding, greater contextual awareness, and interactive feedback mechanisms that enable users to guide or correct the AI's assumptions in real time.

5.4 Design Implications for Human AI Assistive Systems for Real-world Accessibility

Given the rapid evolution of LMMs, the challenges identified above point to several key design implications for improving real-world accessibility in human-AI assistive systems.

1) Personalization and Adaptability in AI Assistive Systems. AI assistive systems should be teachable and adaptive to users' individual abilities [38, 41, 67], which could be addressed through prompt engineering or model fine-tuning. As noted in Sections 4.2.2 and 4.1.3, participants frequently informed ChatGPT of their vision profile or whether they were using a cane to receive more relevant descriptions. Inspired by professional RSA services like Aira [1], which adapt to user profiles and preferences over time [57, 58, 97], future AI systems could similarly allow users to specify their vision profiles (e.g., blind, low vision, colorblind, tunnel vision) and information preferences (e.g., highlighting barriers when without mobility aids) directly through system prompts or persistent memory. This personalization could enable more effective and relevant descriptions by reducing redundant information, especially details users can already perceive through their own abilities or assistive tools, such as detecting terrain with a cane (P4, P6) or identifying people by voice (P6). Users could also fine-tune models with personal data (e.g., images of their belongings) [41, 67] for more personalized experiences.

2) Building Spatial Understanding and Interaction Memory for Proactive and Reliable AI Guidance. AI assistive systems should support spatial understanding and interaction memory across sessions to facilitate deeper contextual awareness (Section 4.2.2), which may necessitate new memory architectures. During the study, we observed that ChatGPT's imprecise spatial understanding often caused confusion and posed potential risks for users. For example, ChatGPT failed to recognize that the user had already turned around, resulting in incorrect and misleading guidance (Section 4.3.3 & 4.1.3). This stands in contrast to professional RSA human agents, who routinely provide environmental information for new scenes to help set expectations for BVI users [43]. Similar limitations in spatial understanding have been observed in prior systems, where spatial information was inferred from isolated images rather than a holistic understanding of the user's surroundings [25]. This gap could be addressed by building spatial understanding from rich visual data [87, 95, 96], potentially in real time, along with the advancement in computing capabilities. Also, given ongoing efforts to build memory for LLM agents [74, 80, 99], such as chat histories for conversational agents, behavioral summaries for simulated human [74], or usage patterns for modeling user intent [80], it is important to investigate appropriate levels of knowledge abstraction for memory architectures in real-world assistive contexts. Based on our observations, contextual cues from user actions could serve as valuable memory resources [52, 53, 55, 56]. For example, P1 had to repeatedly ask the same question while performing the same physical action, such as holding up each spray bottle to the camera (Section 4.1.2). Such actions could be stored as memory cues to prompt the AI to proactively respond without needing repeated speech queries. Future work should explore the development of spatial understanding in AI assistive systems, and effective methods for delivering navigation guidance (e.g., clock directions) informed by trained human guidance [43]. Additionally, building interaction memory could help reduce the user's burden of repeated querying.

3) Balancing Human-Like Perception and Functional Reliability in Assistive AI Systems for Long-Term Adoption The advancement of computer vision, natural language processing, and speech generation is increasingly blurring the boundaries between AI and humans [34, 74]. Correspondingly, our findings in Section 4.3 revealed that beyond traditional text presentation by screen reader, participants perceived ChatGPT as distinctly human-like due to its high-quality voice, natural conversation style and emotionally supportive responses (e.g., consistently expressing consent, a problem commonly recognized as *"sycophancy"* in LMMs [61, 76]). However, unlike human agents in RSA services, who can proactively describe relevant content and express reduced confidence in unfamiliar environments [43], AI systems may adopt a sycophantic and overly confirmatory tone. This can be double-edged, which potentially creates a false sense of security and overstating the reliability of information when AI simulates human interaction without truly matching human-level capabilities [17], such as by misleading users into thinking it can proactively describe dynamic contexts (Section 4.1.2) or provide accurate directions (Section 4.3.3). Such concerns are especially critical in high-stakes, real-world applications involving disabled users, who may be vulnerable to AI errors [11, 36, 39, 44, 91] and misjudgments introduced by over-reliance on AI systems [10, 30, 66].

Choung et al. [27] proposed two models of AI trust: *Human-like trust* refers to the ethical and social qualities embedded in AI such as its perceived transparency and justice [51, 81] while *functionality-based trust* focuses on technical aspects, such as performance and reliability [49, 51]. Although users may be drawn to an AI system's human-like traits, sustained trust and long-term adoption are more strongly influenced by the system's ability to perform reliably and meet expectations [27]. As AI technologies continue to evolve toward greater human-likeness, it becomes increasingly important to manage the tension between perceived realism and actual capability. Users must be supported in forming accurate expectations and discouraged from misattributing uniquely human traits, such as moral reasoning, emotional awareness, or expressive characteristics (e.g., voice, tone), to AI systems. Future work in the design of real-world AI assistive systems should explicitly incorporate trust as a foundational construct, ensuring that the system's perceived humanness (*human-like trust*) is appropriately aligned with its actual capabilities (*functionality-based trust*) for long-term adoption.

5.5 Limitations of Study

At the time of our study, ChatGPT's "Advanced Voice with Video" [3] feature was subscription-based with high monthly costs and subject to daily usage limits. Even though the partnership with Be My Eyes [2] was announced, it remained largely inaccessible to most potential users. Consequently, it was challenging to observe how users naturally integrated this technology into their everyday lives, as explored in previous studies [11, 16, 22, 31, 36, 72, 73, 75, 91, 92]. To address this, we designed various task scenarios informed by prior research to actively engage users with this feature and obtained key findings. Future research should further investigate how BVI users leverage such advanced capabilities through diary or field studies to provide insights into more practical and diverse scenarios encountered in their daily life. Second, while we aimed to recruit participants with diverse visual abilities, geographic and local recruitment constraints limited our sample to six fully blind individuals and two with low vision. Consequently, our findings may not fully capture the broader spectrum of experiences among the low-vision community. Future research should include participants with a wider variety of visual abilities, diverse perspectives, and different assistive tools to better generalize these insights.

6 Conclusion

In this paper, we explored the gaps in LMM-powered live video AI for real-world assistive applications through a user study with BVI participants. Utilizing ChatGPT Advanced Voice with Video, a state-of-the-art live video AI technology as of 2025, participants engaged with various scenarios representative of everyday tasks. Our findings indicated that while such AI systems provide useful support through conversation, they exhibited notable limitations, including difficulties in delivering precise navigational information and providing live descriptions essential for real-world contexts. Also, ChatGPT frequently assumed its users possessed visual capabilities, leading to inappropriate and ineffective responses to user requests. Other major limitations included inaccurate visual and spatial memory, overly affirmative responses, and a reliance on general world knowledge rather than immediate reality. Based on these insights, we discussed implications for developing context-aware, adaptive live video AI agents tailored specifically to real-world assistive needs.

Acknowledgments

We thank our anonymous reviewers and all the participants in our study for their feedback and suggestions. This work was partially funded by an unrestricted gift from Google.

References

[1] 2025. Aira. https://aira.io/
[2] 2025. BeMyEyes. https://www.bemyeyes.com/
[3] 2025. ChatGPT can now see, hear, and speak. https://openai.com/index/chatgpt-can-now-see-hear-and-speak/
[4] 2025. Envision AI. https://www.letsenvision.com/
[5] 2025. Introducing Be My AI (formerly Virtual Volunteer) for People who are Blind or Have Low Vision, Powered by OpenAI's GPT-4. https://www.bemyeyes.com/blog/introducing-be-my-eyes-virtual-volunteer
[6] 2025. Orcam: Empowering Accessibility with AI. https://www.orcam.com/en-us/home?srsltid=AfmBOoqOElsZ58Z5kmCs_LDgPW7eq4WRpLVMbFGtydZn4aqayDxc7wxR
[7] 2025. PiccyBot. https://www.piccybot.com/
[8] 2025. SeeingAI. https://www.seeingai.com/
[9] Dragan Ahmetovic, Cole Gleason, Chengxiong Ruan, Kris Kitani, Hironobu Takagi, and Chieko Asakawa. 2016. NavCog: a navigational cognitive assistant for the blind. In *Proceedings of the 18th international conference on human-computer interaction with mobile devices and services*. 90–99.
[10] Rahaf Alharbi, Robin N. Brewer, and Sarita Schoenebeck. 2022. Understanding Emerging Obfuscation Technologies in Visual Description Services for Blind and Low Vision People. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 469 (Nov. 2022), 33 pages. https://doi.org/10.1145/3555570
[11] Rahaf Alharbi, Pa Lor, Jaylin Herskovitz, Sarita Schoenebeck, and Robin N. Brewer. 2024. Misfitting With AI: How Blind People Verify and Contest AI Errors. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 61, 17 pages. https://doi.org/10.1145/3663548.3675659
[12] Syed Rizal Alfam Wan Alwi and Mohamad Noh Ahmad. 2013. Survey on outdoor navigation system needs for blind people. In *2013 IEEE student conference on research and developement*. IEEE, 144–148.
[13] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233
[14] Aries Arditi, Emily Holmes, Peter Reedijk, and Roger Whitehouse. 1999. Interactive tactile maps, visual disability, and accessibility of building interiors. *Visual Impairment Research* 1, 1 (1999), 11–21.
[15] Masatoshi Arikawa, Shin'ichi Konomi, and Keisuke Ohnishi. 2007. Navitime: Supporting pedestrian navigation in the real world. *IEEE Pervasive Computing* 6, 3 (2007), 21–29.
[16] Mauro Avila, Katrin Wolf, Anke Brock, and Niels Henze. 2016. Remote Assistance for Blind Users in Daily Life: A Survey about Be My Eyes. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (Corfu, Island, Greece) *(PETRA '16)*. Association for Computing Machinery, New York, NY, USA, Article 85, 2 pages. https://doi.org/10.1145/2910674.2935839
[17] Tita Alissa Bach, Amna Khan, Harry Hallock, Gabriela Beltrão, and Sonia Sousa. 2024. A systematic literature review of user trust in AI-enabled systems: An HCI perspective. *International Journal of Human–Computer Interaction* 40, 5 (2024), 1251–1266.
[18] Maryam Bandukda, Aneesha Singh, Nadia Berthouze, and Catherine Holloway. 2019. Understanding experiences of blind individuals in outdoor nature. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.
[19] Oliver Bendel. 2024. How Can Generative AI Enhance the Well-being of Blind?. In *Proceedings of the AAAI Symposium Series*, Vol. 3. 340–347.
[20] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. 2010. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology* (New York, New York, USA) *(UIST '10)*. Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/1866029.1866080
[21] Jeffrey P Bigham, Chandrika Jayant, Andrew Miller, Brandyn White, and Tom Yeh. 2010. VizWiz:: LocateIt-enabling blind people to locate objects in their environment. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 65–72.
[22] Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P. Bigham. 2013. Visual challenges in the everyday lives of blind people. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2117–2126. https://doi.org/10.1145/2470654.2481291
[23] Erin L. Brady, Yu Zhong, Meredith Ringel Morris, and Jeffrey P. Bigham. 2013. Investigating the appropriateness of social network question asking as a resource for blind users. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, Texas, USA) *(CSCW '13)*. Association for Computing Machinery, New York, NY, USA, 1225–1236. https://doi.org/10.1145/2441776.2441915
[24] Michael Brambring. 1985. Mobility and orientation processes of the blind. In *Electronic Spatial Sensing for the Blind: Contributions from Perception, Rehabilitation, and Computer Vision*. Springer, 493–508.
[25] Ruei-Che Chang, Yuxuan Liu, and Anhong Guo. 2024. WorldScribe: Towards Context-Aware Live Visual Descriptions. 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 140, 18 pages. https://doi.org/10.1145/3654777.3676375
[26] Ruei-Che Chang, Yuxuan Liu, Lotus Zhang, and Anhong Guo. 2024. EditScribe: Non-Visual Image Editing with Natural Language Verification Loops. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 65, 19 pages. https://doi.org/10.1145/3663548.3675599
[27] Hyesun Choung, Prabu David, and Arun Ross. 2023. Trust in AI and its role in the acceptance of AI technologies. *International Journal of Human–Computer Interaction* 39, 9 (2023), 1727–1739.
[28] Thomas Gallagher, Elyse Wise, Binghao Li, Andrew G Dempster, Chris Rizos, and Euan Ramsey-Stewart. 2012. Indoor positioning system based on sensor fusion for the blind and visually impaired. In *2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 1–9.
[29] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.
[30] Kate S Glazko, Momona Yamagami, Aashaka Desai, Kelly Avery Mack, Venkatesh Potluri, Xuhai Xu, and Jennifer Mankoff. 2023. An Autoethnographic Case Study of Generative Artificial Intelligence's Utility for Accessibility. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA) *(ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, Article 99, 8 pages. https://doi.org/10.1145/3597638.3614548
[31] Ricardo E Gonzalez Penuela, Jazmin Collins, Cynthia Bennett, and Shiri Azenkot. 2024. Investigating Use Cases of AI-Powered Scene Description Applications for Blind and Low Vision People. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 901, 21 pages. https://doi.org/10.1145/3613904.3642211
[32] Christina Granquist, Susan Y Sun, Sandra R Montezuma, Tu M Tran, Rachel Gage, and Gordon E Legge. 2021. Evaluation and comparison of artificial intelligence vision aids: Orcam myeye 1 and seeing ai. *Journal of Visual Impairment &*

Blindness 115, 4 (2021), 277–285.

[33] João Guerreiro, Daisuke Sato, Saki Asakawa, Huixu Dong, Kris M Kitani, and Chieko Asakawa. 2019. Cabot: Designing and evaluating an autonomous navigation robot for blind people. In *Proceedings of the 21st international ACM SIGACCESS conference on computers and accessibility*. 68–82.

[34] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 433, 19 pages. https://doi.org/10.1145/3544548.3580688

[35] R. Hartson and P.S. Pyla. 2012. *The UX Book: Process and Guidelines for Ensuring a Quality User Experience*. Elsevier Science. https://books.google.ca/books?id=w4I3Y64SWLoC

[36] Jaylin Herskovitz, Andi Xu, Rahaf Alharbi, and Anhong Guo. 2023. Hacking, Switching, Combining: Understanding and Supporting DIY Assistive Technology Design by Blind People. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 57, 17 pages. https://doi.org/10.1145/3544548.3581249

[37] Nicole Holmes and Kelly Prentice. 2015. iPhone video link as an orientation tool: Remote O&M for people with vision impairment. *Vision Rehabilitation International* 7, 1 (2015), 60–67.

[38] Jonggi Hong, Jaina Gandhi, Ernest Essuah Mensah, Farnaz Zamiri Zeraati, Ebrima Jarjue, Kyungjun Lee, and Hernisa Kacorri. 2022. Blind Users Accessing Their Training Images in Teachable Object Recognizers. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 14, 18 pages. https://doi.org/10.1145/3517428.3544824

[39] Jonggi Hong and Hernisa Kacorri. 2024. Understanding How Blind Users Handle Object Recognition Errors: Strategies and Challenges. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 63, 15 pages. https://doi.org/10.1145/3663548.3675635

[40] Karst MP Hoogsteen, Sarit Szpiro, Gabriel Kreiman, and Eli Peli. 2022. Beyond the cane: describing urban scenes to blind people for mobility tasks. *ACM Transactions on Accessible Computing (TACCESS)* 15, 3 (2022), 1–29.

[41] Hernisa Kacorri, Kris M. Kitani, Jeffrey P. Bigham, and Chieko Asakawa. 2017. People with Visual Impairment Training Personal Object Recognizers: Feasibility and Challenges. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5839–5849. https://doi.org/10.1145/3025453.3025899

[42] Amy A Kalia, Gordon E Legge, Rudrava Roy, and Advait Ogale. 2010. Assessment of indoor route-finding technology for people who are visually impaired. *Journal of visual impairment & blindness* 104, 3 (2010), 135–147.

[43] Rie Kamikubo, Naoya Kato, Keita Higuchi, Ryo Yonetani, and Yoichi Sato. 2020. Support Strategies for Remote Guides in Assisting People with Visual Impairments for Effective Indoor Navigation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376823

[44] Hyung Nam Kim. 2022. User experience of assistive apps among people with visual impairment. *Technology and Disability* 34, 3 (2022), 165–174.

[45] Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022. Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics. *arXiv preprint arXiv:2205.10646* (2022).

[46] Vladimir Kulyukin, Chaitanya Gharpure, John Nicholson, and Grayson Osborne. 2006. Robot-assisted wayfinding for the visually impaired in structured indoor environments. *Autonomous robots* 21 (2006), 29–41.

[47] Elizabeth Kupferstein, Yuhang Zhao, Shiri Azenkot, and Hathaitorn Rojnirun. 2020. Understanding the use of artificial intelligence based visual aids for people with visual impairments. *Investigative Ophthalmology & Visual Science* 61, 7 (2020), 932–932.

[48] Masaki Kuribayashi, Kohei Uehara, Allan Wang, Shigeo Morishima, and Chieko Asakawa. 2025. WanderGuide: Indoor Map-less Robotic Guide for Exploration by Blind People. *arXiv preprint arXiv:2502.08906* (2025).

[49] Nancy K Lankton, D Harrison McKnight, and John Tripp. 2015. Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems* 16, 10 (2015), 1.

[50] Jaewook Lee, Jaylin Herskovitz, Yi-Hao Peng, and Anhong Guo. 2022. ImageExplorer: Multi-Layered Touch Exploration to Encourage Skepticism Towards Imperfect AI-Generated Image Captions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 462, 15 pages. https://doi.org/10.1145/3491102.3501966

[51] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[52] Kyungjun Lee, Jonggi Hong, Simone Pimento, Ebrima Jarjue, and Hernisa Kacorri. 2019. Revisiting Blind Photography in the Context of Teachable Object Recognizers. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 83–95. https://doi.org/10.1145/3308561.3353799

[53] Kyungjun Lee and Hernisa Kacorri. 2019. Hands Holding Clues for Object Recognition in Teachable Machines. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300566

[54] Kyungyeon Lee, Sohyeon Park, and Uran Oh. 2021. Designing Product Descriptions for Supporting Independent Grocery Shopping of People with Visual Impairments. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 425, 6 pages. https://doi.org/10.1145/3411763.3451806

[55] Kyungjun Lee, Abhinav Shrivastava, and Hernisa Kacorri. 2020. Hand-priming in object localization for assistive egocentric vision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3422–3432.

[56] Kyungjun Lee, Abhinav Shrivastava, and Hernisa Kacorri. 2021. Leveraging hand-object interactions in assistive egocentric vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 6 (2021), 6820–6831.

[57] Sooyeon Lee, Madison Reddie, Krish Gurdasani, Xiying Wang, Jordan Beck, Mary Beth Rosson, and John M Carroll. 2018. Conversations for Vision: Remote Sighted Assistants Helping People with Visual Impairments. *arXiv preprint arXiv:1812.00148* (2018).

[58] Sooyeon Lee, Madison Reddie, Chun-Hua Tsai, Jordan Beck, Mary Beth Rosson, and John M. Carroll. 2020. The Emerging Professional Practice of Remote Sighted Assistance for People with Visual Impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376591

[59] Yue Lian, De-er Liu, and Wei-zhen Ji. 2024. Survey and analysis of the current status of research in the field of outdoor navigation for the blind. *Disability and Rehabilitation: Assistive Technology* 19, 4 (2024), 1657–1675.

[60] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[61] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253* (2024).

[62] Xingyu Bruce Liu, Shitao Fang, Weiyan Shi, Chien-Sheng Wu, Takeo Igarashi, and Xiang 'Anthony' Chen. 2025. Proactive Conversational Agents with Inner Thoughts. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 184, 19 pages. https://doi.org/10.1145/3706598.3713760

[63] Zhe Liu, Chunyang Chen, Junjie Wang, Mengzhuo Chen, Boyu Wu, Yuekai Huang, Jun Hu, and Qing Wang. 2024. Unblind Text Inputs: Predicting Hint-text of Text Input in Mobile Apps via LLM. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 51, 20 pages. https://doi.org/10.1145/3613904.3642939

[64] Jack M Loomis, Reginald G Golledge, and Roberta L Klatzky. 1998. Navigation system for the blind: Auditory display modes and guidance. *Presence* 7, 2 (1998), 193–203.

[65] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[66] Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding Blind People's Experiences with Computer-Generated Captions of Social Media Images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5988–5999. https://doi.org/10.1145/3025453.3025814

[67] Cecily Morrison, Martin Grayson, Rita Faia Marques, Daniela Massiceti, Camilla Longden, Linda Wen, and Edward Cutrell. 2023. Understanding Personalized Accessibility through Teachable AI: Designing and Evaluating Find My Things for People who are Blind or Low Vision. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA) *(ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, Article 31, 12 pages. https://doi.org/10.1145/3597638.3608395

[68] Annika Muehlbradt and Shaun K Kane. 2022. What's in an ALT Tag? Exploring Caption Content Priorities through Collaborative Captioning. *ACM Transactions*

on Accessible Computing (TACCESS) 15, 1 (2022), 1–32.

[69] Masayuki Murata, Dragan Ahmetovic, Daisuke Sato, Hironobu Takagi, Kris M Kitani, and Chieko Asakawa. 2018. Smartphone-based indoor localization for blind navigation across building complexes. In *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 1–10.

[70] Jongwhoa Na. 2006. The blind interactive guide system using RFID-based indoor positioning system. In *International Conference on Computers for Handicapped Persons*. Springer, 1298–1305.

[71] Vishnu Nair, Hanxiu 'Hazel' Zhu, and Brian A. Smith. 2023. ImageAssist: Tools for Enhancing Touchscreen-Based Image Exploration Systems for Blind and Low Vision Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 76, 17 pages. https://doi.org/10.1145/3544548.3581302

[72] Brian J Nguyen, William S Chen, Allison J Chen, Andrew Utt, Emily Hill, Ryan Apgar, and Daniel L Chao. 2019. Large-scale assessment of needs in low vision individuals using the Aira assistive technology. *Clinical ophthalmology* (2019), 1853–1868.

[73] Brian J Nguyen, Yeji Kim, Kathryn Park, Allison J Chen, Scarlett Chen, Donald Van Fossan, and Daniel L Chao. 2018. Improvement in patient-reported quality of life outcomes in severely visually impaired individuals using the Aira assistive technology system. *Translational vision science & technology* 7, 5 (2018), 30–30.

[74] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 2, 22 pages. https://doi.org/10.1145/3586183.3606763

[75] Ricardo E Gonzalez Penuela, Ruiying Hu, Sharon Lin, Tanisha Shende, and Shiri Azenkot. 2025. Towards Understanding the Use of MLLM-Enabled Applications for Visual Interpretation by Blind and Low Vision People. *arXiv preprint arXiv:2503.05899* (2025).

[76] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*. 13387–13434.

[77] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173677

[78] Lisa Ran, Sumi Helal, and Steve Moore. 2004. Drishti: an integrated indoor/outdoor blind navigation system and service. In *Second IEEE annual conference on pervasive computing and communications, 2004. Proceedings of the*. IEEE, 23–30.

[79] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[80] Omar Shaikh, Shardul Sapkota, Shan Rizvi, Eric Horvitz, Joon Sung Park, Diyi Yang, and Michael S Bernstein. 2025. Creating General User Models from Computer Use. *arXiv preprint arXiv:2505.10831* (2025).

[81] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International journal of human-computer studies* 146 (2021), 102551.

[82] Shraga Shoval, Johann Borenstein, and Yoram Koren. 1994. Mobile robot obstacle avoidance in a computerized travel aid for the blind. In *Proceedings of the 1994 IEEE International Conference on Robotics and Automation*. IEEE, 2023–2028.

[83] Abigale Stangl, Shasta Ihorn, Yue-Ting Siu, Aditya Bodi, Mar Castanon, Lothar D Narins, and Ilmi Yoon. 2023. The Potential of a Visual Dialogue Agent In a Tandem Automated Audio Description System for Videos. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA) *(ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, Article 32, 17 pages. https://doi.org/10.1145/3597638.3608402

[84] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376404

[85] Abigale Stangl, Nitin Verma, Kenneth R Fleischmann, Meredith Ringel Morris, and Danna Gurari. 2021. Going beyond one-size-fits-all image descriptions to satisfy the information wants of people who are blind or have low vision. In *Proceedings of the 23rd international ACM SIGACCESS conference on computers and accessibility*. 1–15.

[86] Yu-Yun Tseng, Alexander Bell, and Danna Gurari. 2022. Vizwiz-fewshot: Locating objects in images taken by people with visual impairments. In *European Conference on Computer Vision*. Springer, 575–591.

[87] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. 2025. VGGT: Visual Geometry Grounded Transformer. *arXiv preprint arXiv:2503.11651* (2025).

[88] Elyse Wise, Binghao Li, Thomas Gallagher, Andrew G Dempster, Chris Rizos, Euan Ramsey-Stewart, and Daniel Woo. 2012. Indoor navigation for the blind and vision impaired: Where are we and where are we going?. In *2012 international conference on Indoor positioning and indoor navigation (IPIN)*. IEEE, 1–7.

[89] Jacob O. Wobbrock, Shaun K. Kane, Krzysztof Z. Gajos, Susumu Harada, and Jon Froehlich. 2011. Ability-Based Design: Concept, Principles and Examples. *ACM Trans. Access. Comput.* 3, 3, Article 9 (April 2011), 27 pages. https://doi.org/10.1145/1952383.1952384

[90] Farrah Wong, R Nagarajan, and Sazali Yaacob. 2003. Application of stereovision in a navigation aid for blind people. In *Fourth international conference on information, communications and signal processing, 2003 and the fourth Pacific Rim conference on multimedia. Proceedings of the 2003 joint*, Vol. 2. IEEE, 734–737.

[91] Jingyi Xie, Rui Yu, He Zhang, Syed Masum Billah, Sooyeon Lee, and John M Carroll. 2025. Beyond Visual Perception: Insights from Smartphone Interaction of Visually Impaired Users with Large Multimodal Models. *arXiv preprint arXiv:2502.16098* (2025).

[92] Andi Xu, Minyu Cai, Dier Hou, Ruei-Che Chang, and Anhong Guo. 2024. ImageExplorer Deployment: Understanding Text-Based and Touch-Based Image Exploration in the Wild *(W4A '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3677846.3677861

[93] Shuchang Xu, Chang Chen, Zichen Liu, Xiaofu Jin, Lin-Ping Yuan, Yukang Yan, and Huamin Qu. 2024. Memory Reviver: Supporting Photo-Collection Reminiscence for People with Visual Impairment via a Proactive Chatbot. 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 88, 17 pages. https://doi.org/10.1145/3654777.3676336

[94] Bufang Yang, Lixing He, Kaiwei Liu, and Zhenyu Yan. 2024. Viassist: Adapting multi-modal large language models for users with visual impairments. In *2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys)*. IEEE, 32–37.

[95] Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. 2025. Fast3R: Towards 3D Reconstruction of 1000+ Images in One Forward Pass. *arXiv preprint arXiv:2501.13928* (2025).

[96] Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 2024. 3D-Mem: 3D Scene Memory for Embodied Exploration and Reasoning. arXiv:2411.17735 [cs.CV] https://arxiv.org/abs/2411.17735

[97] Rui Yu, Sooyeon Lee, Jingyi Xie, Syed Masum Billah, and John M Carroll. 2024. Human–AI collaboration for remote sighted assistance: Perspectives from the LLM era. *Future internet* 16, 7 (2024), 254.

[98] He Zhang, Nicholas J. Falletta, Jingyi Xie, Rui Yu, Sooyeon Lee, Syed Masum Billah, and John M. Carroll. 2025. Enhancing the Travel Experience for People with Visual Impairments through Multimodal Interaction: NaviGPT, A Real-Time AI-Driven Mobile Navigation System. In *Companion Proceedings of the 2025 ACM International Conference on Supporting Group Work* (Hilton Head, New Jersey, USA) *(GROUP '25)*. Association for Computing Machinery, New York, NY, USA, 29–35. https://doi.org/10.1145/3688828.3699636

[99] Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501* (2024).

[100] Yu Zhong, Walter S Lasecki, Erin Brady, and Jeffrey P Bigham. 2015. Regionspeak: Quick comprehensive spatial descriptions of complex images for blind users. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2353–2362.