Enabling Real-World Assistive Agents: From Live Vision to Proactive Context-Aware Information Delivery

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

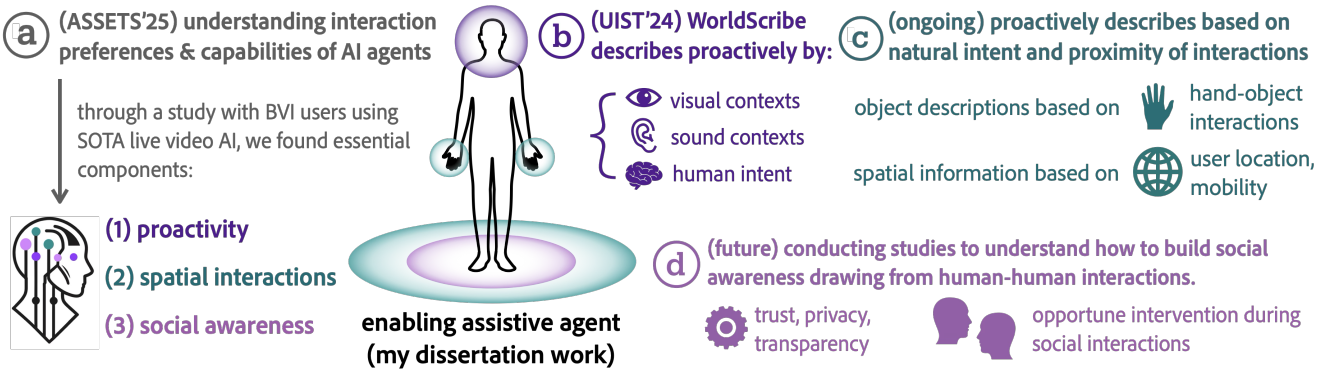

Figure 1: My dissertation work explores the design of an assistive agent as an always-available companion to assist BVI people in real-world interactions. (a) I first conducted a user study with eight BVI people to probe the design insights for designing assistive agents [11]. Based on the insights, (b) I developed a system, WorldScribe, that proactively provides live visual descriptions based on user intent, and visual and sound contexts [10]. (c) My ongoing works expand WorldScribe to take natural user interactions to provide corresponding information for BVI people, such as object descriptions for hand-object interactions, and spatial information based on location and mobility. (d) In the future, I aim to draw insights from human assistance to develop an assistive agent that serves as a long-term, context-aware companion in everyday environments.

Abstract

Interacting with the real world is a fundamental part of daily life, yet it remains challenging for individuals who are blind or visually impaired (BVI). It demands live, contextual understanding of dynamic environments, along with interactive, multimodal communication to fulfill their sensory and cognitive needs. To address this, my dissertation develops assistive AI systems and frameworks that observe the real world through multimodal sensing, reason about essential information in response to user contexts, and deliver human-like verbal communication to support real-world understanding. First, I explored the design insights of assistive AI agents by investigating how BVI users interact with human-like video AI systems across diverse real-world contexts. Second, with the identified insights, such as a lack of proactivity, I developed a mobile application that analyzes live camera feeds to generate real-time visual descriptions aligned with user goals, delivering them in harmony with the audio environment. Lastly, I extend it with a set of human-like capabilities, such as memory, spatial understanding, and the ability to infer intent from natural interactions, to act as a long-term assistive companion. Ultimately, my dissertation advances a paradigm shift from digital agents to real-world assistive agents that enhance the independence and agency of BVI individuals.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; **Accessibility technologies**.

Keywords

Visual descriptions, blind, visually impaired, assistive technology, accessibility, context-aware, customization, LLM, real world, sound

ACM Reference Format:
Ruei-Che Chang. 2025. Enabling Real-World Assistive Agents: From Live Vision to Proactive Context-Aware Information Delivery. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct '25), September 28–October 01, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3746058.3758468

1 Introduction

Interacting with the real world is a fundamental part of daily life, yet everyday tasks, such as grocery shopping [15, 17, 18, 21], cooking [19, 20], physical making [12, 13], and navigating unfamiliar spaces [16], remain challenging for individuals who are blind or visually impaired (BVI). Unlike digital tasks and interfaces that have rich metadata for promoting accessibility, these physical activities demand real-time contextual understanding and effective

UIST Adjunct '25, Busan, Republic of Korea

ACM ISBN 979-8-4007-2036-9/25/09
https://doi.org/10.1145/3746058.3758468

communication to address individuals' sensory and cognitive needs. Current mobile applications, such as SeeingAI [4] and BeMyAI [3], enable BVI users to snap a photo, receive visual descriptions within seconds, and query missing details through conversation. Yet, their utility falls short in the rapidly changing real world that requires live context-aware descriptions. Remote sighted assistance (RSA) connects BVI users with sighted agents who provide support through video calls and real-time verbal guidance. However, RSA services are not always available, costly, and raise privacy concerns [8].

The recent rise of agentic interactions, AI systems capable of perceiving, reasoning, and acting over time to complete digital tasks, has primarily occurred in well-structured, digital environments rich in metadata. In contrast, real-world settings relevant to BVI users are dynamic, unstructured, and individually different. How to capture and represent meaningful contextual information from these environments in ways that support agentic AI aligned with BVI users' needs remains an open challenge.

To address these questions, my research develops assistive AI systems and frameworks that observe the real world through multimodal sensing, reason about essential information in response to user contexts, and use human-like verbal communication to support decision-making. I envision an AI companion that is always available to BVI users and remains aware of both their past and present surroundings. It provides timely, contextually useful suggestions while respecting social norms and privacy, simulating the presence of a trusted sighted companion to support the independence and agency of BVI individuals.

To achieve this goal, I outline the three stages of my research agenda in the following sections:

(i) Conducting user studies with BVI users **to understand the interaction needs and preferences of BVI users with assistive AI agents** across diverse real-world contexts [11].

(ii) Developing an assistive AI system, WorldScribe [10], that provides context-aware live descriptions **to empower BVI users with timely, relevant information for understanding the real world.**

(iii) Extending it with a set of human-like capabilities, such as memory, spatial understanding, observing social cues, and the ability to infer intent from natural tactile interactions, **to simulate a long-term, context-aware companion in everyday environments.**

Ultimately, my dissertation will contribute to a real-world assistive agent that is context-aware, socially and environmentally adaptive, and capable of providing always-available, trustworthy support for BVI individuals in dynamic everyday settings. My aim is to advance a paradigm shift from digital agents to embodied, real-world assistive agents that meaningfully enhance the independence and agency of BVI individuals.

2 Understanding BVI Interaction Needs and Preferences with Live Video AI System (ASSETS'25)

Recent advancements in large multimodal models (LMMs) have advanced assistive technologies, enabling BVI individuals to access their environments more independently. Traditionally, users relied on sighted assistance, either by sending photos to crowd workers for answers [7, 24] or by using real-time remote services like BeMyEyes [2] and Aira [1]. In recent years, LMM-powered applications such as Be My AI [3] and SeeingAI [4] have further expanded autonomy by supporting AI-driven conversations around static images. However, there are still gaps between such LLM-powered applications and human assistance, which can observe complex visual and audio context in real time and provide useful descriptions according to user needs and different abilities.

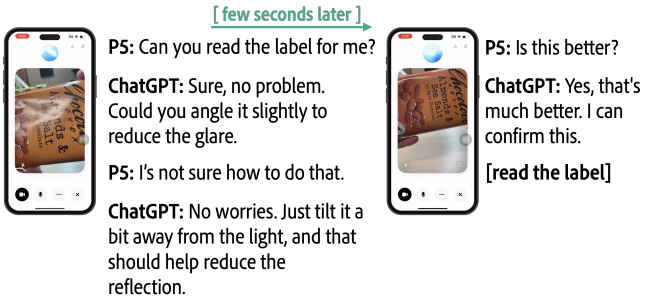

Figure 2: Illustration of user turn-taking interactions with ChatGPT's live video AI. The screenshots were cropped from the video recording of a participant in our study [11].

Advancing this trajectory, OpenAI released ChatGPT Advanced Voice with Video on December 12, 2024 [5], a state-of-the-art system enabling visual question answering (VQA) through live video. This development marked a notable shift from static image interaction to real-time engagement with dynamic physical environments. While prior research has begun to examine how BVI users adopt these emerging tools in everyday life [6, 14, 22, 23, 25, 26], important questions remain about how live video AI systems perform in realistic, task-oriented contexts.

To address this gap, we conducted an in-person user study with eight BVI participants (*To appear at ASSETS'25* [11]), exploring a range of real-world task scenarios designed to engage them with ChatGPT's live video AI [5]. These scenarios included understanding unknown objects, distinguishing between similar items, recognizing prominent visual landmarks, and locating specific targets within unfamiliar indoor and outdoor environments. Our findings suggest that while ChatGPT demonstrated promising capabilities of describing visual content, notable limitations remained across several dimensions that impact their effectiveness as assistive tools for BVI individuals. We highlight the following key challenges:

(i) **ChatGPT lacks proactivity and the ability to continuously describe essential visual information**, even when explicitly requested by users in scenarios such as identifying obstacles or locating specific objects as they move. Instead, it responded only to direct prompts, requiring users to repeatedly ask similar questions, which caused frustration and reduced efficiency, especially during dynamic or repetitive tasks.

(ii) **ChatGPT lacks spatial perception and memory for accurate directional guidance**, leading to confusion and inefficiency during tasks such as object search and outdoor navigation, where recalling prior movement is essential. BVI users demand descriptions to complement their orientation and mobility (O&M) skills, such as obstacles, signs, and landmarks, especially in situations where auditory cues or residual vision are limited.

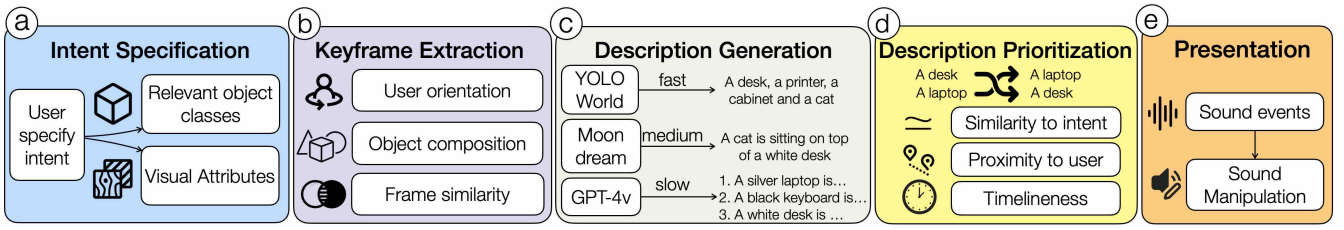

Figure 3: WorldScribe system architecture. (a) The user first specifies their intent through speech and WorldScribe decomposes it into specific visual attributes and relevant objects. (b) WorldScribe extracts keyframes based on user orientation, object compositions, and frame similarity. (c) Next, it generates candidate descriptions with a suite of visual and language models. (d) WorldScribe then prioritizes the descriptions based on the user's intent, proximity to the user, and relevance to the current visual context. (e) Finally, it detects environmental sounds and manipulates the presentation of the descriptions accordingly.

(iii) **ChatGPT fosters misplaced trust in its abilities through affirming responses, sycophancy, and natural conversation.** These human-like traits can inadvertently create a false sense of security and inflate users' perception of information reliability, a concern that is particularly critical in high-stakes assistive contexts, where overreliance on AI may lead to harmful outcomes.

To address these three challenges, my next steps are to develop a live description system with proactivity that provides essential information, and a set of human-like capabilities toward realizing a real-world assistive agent.

3 Context-Aware Live Descriptions for BVI Navigation and Interaction (UIST'24 & DIS'24)

To enable **proactivity and live descriptions**, I developed WorldScribe [10], a system that delivers live visual descriptions based on users' explicit intent (e.g., stated goals) and implicit intent (e.g., inferred from camera motion).

WorldScribe achieves this by interpreting camera motion as an information cue: when the user moves the camera, it provides brief, high-level descriptions, while remaining stationary triggers more detailed ones. These dynamic descriptions are powered by a suite of vision-language models (VLMs) designed to balance richness and latency to enable real-time performance. Additionally, building on my prior work, SoundShift [9] in DIS '24, which explores sound manipulation techniques to enhance the distinction of virtual sounds in real-world environments, WorldScribe incorporates sound context to modulate description delivery in different ways, such as increasing its volume in noisy environments or pausing when speech is detected. The descriptions were also prioritized based on user-defined intent and the proximity of the described content (e.g., barriers closer to the user). BVI users can also customize their desired visual information, its granularity, and audio presentation (Figure 5).

An evaluation with six BLV people shows that they consider WorldScribe descriptions useful with adaptive and customized visual information in different contexts. They also perceive descriptions to be accurate based on the clues they ascertained (e.g., showing their hands within the camera and getting descriptions accordingly). A subsequent technical evaluation shows that the WorldScribe descriptions achieve an accuracy of (84.26%), cover 75% essential user-desired information, and successfully prioritize 80.83% descriptions based on user intent and proximity of the described content.

Through this work, I learn that building an assistive agent is a long-tail challenge, given the wide diversity of users' needs and the complexity and variability of real-world environments. Below, I describe my ongoing and future work toward this goal.

4 Towards An Assistive Agent in the Real World: Expanding WorldScribe to Natural User Interactions (Ongoing & Future Work)

Building on the insights from the above works, I present three ongoing and future directions aimed at realizing a real-world assistive agent. First, expanding WorldScribe's use of camera motion for visual information access, I am exploring how natural hand-object interactions can serve as intent cues for delivering live descriptions within physical reach. Second, as identified in our ASSETS '25 study that spatial awareness and memory are critical for real-world understanding, I am designing a system architecture to generate real-time spatial descriptions, such as direction, distance, object size, and spatial changes. Third, beyond providing real-world information, I aim to explore how to present it in ways that foster trust and reduce human-AI friction.

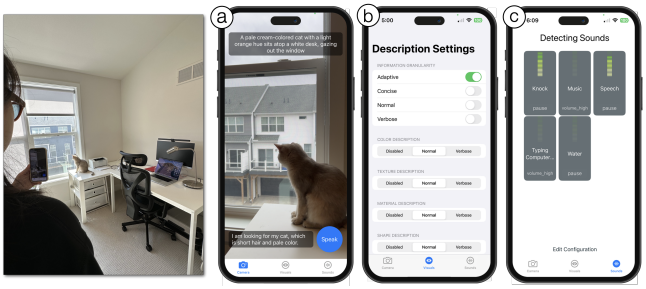

Figure 4: WorldScribe [10] is a system that proactively describes the real world based on user contexts. (a) The user can specify their intent and needs regarding visual attributes or audio presentation through speech input. (b) Besides speech, they can manually select options for richness and other visual attributes. (c) They can also configure pauses or increase the volume of descriptions if certain sound events are detected.

Systems	Context References: determine timing to describe proactively	Description Types: Prioritize rich information based on different contexts
WorldScribe (UIST'24)	**sensory contexts:** visual changes from camera motion and audio events	**visual descriptions:** object labels, scene overviews, hierarchical visual detail
WorldScribe:Touch (manuscript ready)	**hand-object interactions:** postures, gestures, new objects	**object attributes:** visuals, color, texts, visual comparison across items
WorldScribe:Space (ongoing)	**mobility and environment:** mobility aids, orientation, footprint	**spatial information:** directions, object sizes, proximity, layout relative to user

Figure 5: Overview of WorldScribe and its extensions with different context-aware description strategies.

Manuscript Ready for Submission: Exploring Natural Hand-Object Interactions as Intent Cues for Live Descriptions. Expanding WorldScribe's use of camera motion for visual information access, my project, WorldScribe::Touch, explores how natural hand-object interactions can serve as intent cues for delivering live descriptions within hand reach.

I first synthesized a set of gestures, combining both those commonly used by sighted individuals and gestures specific to BVI users. These were informed by prior research on on-device interactions and discreet on-body gestures, including actions like grab, touch, point, and swipe up. Building on this gesture set, I trained lightweight recognizers capable of detecting these gestures and providing live descriptions to describe essential object information in real time. Specifically, WorldScribe::Touch provides immediate feedback when certain gestures are detected: for example, describing an object's attributes when it is touched or grabbed, comparing visual differences between two objects when held, or reading out text when the user performs a swipe-up gesture on the object, mirroring the mobile screen reader interactions when retrieving texts.

We evaluated WorldScribe::Touch with eight BVI participants across a range of object understanding tasks. Participants commended the system for the completeness and accuracy of its descriptions, as well as the intuitive gesture-based method for accessing specific details, especially when compared to the assistive tools they currently use that require frequent photo-taking and query.

Ongoing Work: Toward Live and Meaningful Spatial Understanding via Smartphone-Based Multi-Modal Sensing Building on findings from our ASSETS '25 study [11] that highlight the importance of spatial awareness and memory for real-world navigation, this project aims to develop an end-to-end solution using accessible smartphone data and LMMs to deliver a broad range of spatial information, in order to address the limitations of prior work that typically supports only a few spatial cues through specialized hardware or software. The system is designed to generate real-time spatial descriptions, such as direction, distance, object size, and spatial changes, by leveraging smartphone inputs like camera pose, RGB-D camera frames, and GPS. These are used to construct structured prompts dynamically to enable accurate, live spatial understanding with persistent environmental memory. We will evaluate the system with BVI users and use the resulting data to create open benchmarks for the HCI and AI communities.

Future Work: Designing Socially Aware Assistive Agents Informed by Human Guidance The systems developed thus far leverage a rich combination of smartphone-accessible signals, including explicit verbal queries, camera motion, and hand–object interactions, to generate real-time visual and spatial descriptions. However, a critical open question remains: *when* is it socially appropriate to present such descriptions? Real-world contexts, particularly those involving face-to-face interactions, introduce complex design challenges. For example, how should an assistive agent convey an interlocutor's facial expressions or body language when a BVI user is actively engaged in conversation? *When is it worth interrupting to provide essential information, and how can such interruptions be delivered in a way that is both respectful and meaningful?* As the next step, I will design and conduct a user study to investigate how users perceive and respond to different styles of agent behavior across varied social contexts, such as comparing command-like (e.g., screen reader) versus human-like agent responses in different social settings, such as one-on-one conversations versus group interactions. Ultimately, my goal is to develop a unified assistive agent that is socially aware, proactive, and effective in everyday settings.

5 Expected Impacts

Recent AI integrations in downstream applications have demonstrated the promise of agents capable of autonomously observing, reasoning, and acting based on user context (e.g., programming, computer use). My project explores how such capabilities can be transferred to real-world practical tasks through accessible multimodal data (e.g., live video feed, microphone). The technical core lies in contextual understanding, enabling systems to offer proactive descriptions or actions. This approach has the potential to advance not only traditional assistive technologies but also a wide range of emerging applications, including AI-driven tourism and hands-free or spatial interactions with intelligent agents.

6 Dissertation Status & Aims For UIST Doctoral Symposium

My dissertation goal is to develop assistive agents that can understand user contexts and needs, and proactively provide essential information to BVI people. As such, the UIST DS would offer valuable feedback on system design and overall framing. I have just completed my 3rd year of the PhD program and plan to propose my dissertation in 2026, with the defense to follow by 2027. Therefore, I look forward to receiving targeted feedback and seeking collaboration and mentorship on my ongoing and future works.

Acknowledgments

I am grateful to my advisor, Prof. Anhong Guo, and collaborators for their generous support and insightful feedback. I also thank the participants in my studies, who shared their experiences and offered valuable insights. Also, I appreciate the research community for their constructive feedback on my publications and career.

References

[1] 2024. Aira. https://aira.io/

[2] 2024. BeMyEyes. https://www.bemyeyes.com/

[3] 2024. 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

[4] 2024. SeeingAI. https://www.seeingai.com/

[5] 2025. ChatGPT can now see, hear, and speak. https://openai.com/index/chatgpt-can-now-see-hear-and-speak/

[6] 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. doi:10.1145/2910674.2935839

[7] 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. doi:10.1145/1866029.1866080

[8] 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. doi:10.1145/2441776.2441915

[9] Ruei-Che Chang, Chia-Sheng Hung, Bing-Yu Chen, Dhruv Jain, and Anhong Guo. 2024. SoundShift: Exploring Sound Manipulations for Accessible Mixed-Reality Awareness. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (IT University of Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 116–132. doi:10.1145/3643834.3661556

[10] 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. doi:10.1145/3654777.3676375

[11] 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 *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility* (Denver, Colorado, USA) *(ASSETS '25)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3663547.3746319

[12] Ruei-Che Chang, Chih-An Tsao, Fang-Ying Liao, Seraphina Yong, Tom Yeh, and Bing-Yu Chen. 2021. Daedalus in the Dark: Designing for Non-Visual Accessible Construction of Laser-Cut Architecture. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 344–358. doi:10.1145/3472749.3474754

[13] Ruei-Che Chang, Wen-Ping Wang, Chi-Huan Chiang, Te-Yen Wu, Zheer Xu, Justin Luo, Bing-Yu Chen, and Xing-Dong Yang. 2021. AccessibleCircuits: Adaptive Add-On Circuit Components for People with Blindness or Low Vision. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 670, 14 pages. doi:10.1145/3411764.3445690

[14] 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. doi:10.1145/3544548.3581249

[15] Vladimir Kulyukin, John Nicholson, and Daniel Coster. 2008. Shoptalk: toward independent shopping by people with visual impairments. In *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility* (Halifax, Nova Scotia, Canada) *(Assets '08)*. Association for Computing Machinery, New York, NY, USA, 241–242. doi:10.1145/1414471.1414518

[16] Masaki Kuribayashi, Tatsuya Ishihara, Daisuke Sato, Jayakorn Vongkulbhisal, Karnik Ram, Seita Kayukawa, Hironobu Takagi, Shigeo Morishima, and Chieko Asakawa. 2023. PathFinder: Designing a Map-less Navigation System for Blind People in Unfamiliar Buildings. 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 41, 16 pages. doi:10.1145/3544548.3580687

[17] Jihyun Lee, Jinsol Kim, and Hyunggu Jung. 2020. Challenges and Design Opportunities for Easy, Economical, and Accessible Offline Shoppers with Visual Impairments. In *Proceedings of the 2020 Symposium on Emerging Research from Asia and on Asian Contexts and Cultures* (Honolulu, HI, USA) *(AsianCHI '20)*. Association for Computing Machinery, New York, NY, USA, 69–72. doi:10.1145/3391203.3391223

[18] 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. doi:10.1145/3411763.3451806

[19] Franklin Mingzhe Li, Michael Xieyang Liu, Shaun K. Kane, and Patrick Carrington. 2024. A Contextual Inquiry of People with Vision Impairments in Cooking. 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 38, 14 pages. doi:10.1145/3613904.3642233

[20] Franklin Mingzhe Li, Ashley Wang, Patrick Carrington, and Shaun K. Kane. 2024. A Recipe for Success? Exploring Strategies for Improving Non-Visual Access to Cooking Instructions. 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 26, 15 pages. doi:10.1145/3663548.3675662

[21] Diego López-de Ipiña, Tania Lorido, and Unai López. 2011. Blindshopping: enabling accessible shopping for visually impaired people through mobile technologies. In *Toward Useful Services for Elderly and People with Disabilities: 9th International Conference on Smart Homes and Health Telematics, ICOST 2011, Montreal, Canada, June 20-22, 2011. Proceedings 9*. Springer, 266–270.

[22] 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.

[23] 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).

[24] 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.

[25] 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).

[26] 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. doi:10.1145/3677846.3677861