

ShelfScout: An AI-Driven Approach to Enhance Shopping Accessibility for Blind and Visually Impaired Individuals

Mansi Dhanania

The Department of Electrical and Computer Engineering
McGill University
Montreal, Quebec, Canada

July 2026



A thesis submitted to McGill University in partial
fulfillment of the requirements of the degree of
Master of Science

©Mansi Dhanania, 2026

Contents

Abstract	viii
Abrégé	x
Acknowledgements	xii
Generative AI Statement	xiii
1 Introduction	1
1.1 ShelfScout	3
1.2 Thesis Structure	4
1.3 Contributions of Collaborators	5
2 Literature Review	6
2.1 BVI Accessibility	7
2.1.1 Evolution of Assistive Tools for BVI	7
2.1.2 Prior Work in BVI Shopping Assistance	9
2.1.3 Unmet Needs in BVI Grocery Shopping	10
2.2 Language Models and Hallucinations	12
2.3 BVI Guidance	15
3 Preliminary System Design	18
3.1 Design Goals	18
3.2 SLIV Framework	19
3.3 Hardware Form-Factor	20
3.4 User Experience	21
3.5 Architecture	22
3.5.1 Intent-Based Classification	23

3.5.2	Visual Reasoning Workflows	25
3.5.3	Summarizer	28
4	Preliminary User Study	29
4.1	Baselines for Comparison	30
4.2	Experimental Setup	31
4.2.1	Aisles	31
4.2.2	Tasks and Prompts	32
4.3	Remote Study Design	34
4.3.1	Study Protocol	34
4.3.2	Ethics and Recruitment	36
4.4	Study Results	37
4.4.1	Researcher-Led Evaluation	38
4.4.2	Participant Feedback	40
4.4.3	Limitations of the Study	44
5	System Evolution	45
5.1	Architectural Changes	45
5.1.1	Intent Classification	46
5.1.2	Object Presence Check	47
5.1.3	Visual Reasoning Workflows	48
5.1.4	Summarizer	53
5.1.5	Fallbacks	54
5.2	System Evaluation	54
6	Conclusion	56
6.1	Discussion and Conclusion	56
6.2	Ongoing Developments	57
6.3	Future Work	58
6.4	Broader Implications	59
6.4.1	Hallucinations	59
6.4.2	Accountability	59
6.4.3	Dependence vs. Autonomy	60

A	Appendix	72
A.1	Recruitment Form	72
A.2	Consent Form	73
A.3	Introductory Script	77
A.4	Pre-Test Questionnaire	80
A.5	Post-Test Questionnaire	84
A.6	Post-Study Semi-Structured Interview	87

List of Figures

3.1	Routing Pipelines for ShelfScout	26
4.1	Remote Study Participant Feedback	41
5.1	Updated Routing Pipelines for ShelfScout	46

List of Tables

3.1	SLIV Task Decomposition	21
3.2	Intent Categories, their corresponding SLIV tasks and triggers.	24
4.1	Images used for evaluation across grocery store sections with their corresponding tasks and prompts.	33
4.2	Summary of hallucination avoidance for each system across grocery store sections and tasks. A check mark (✓) indicates that the system avoided hallucinating visual elements in its response. A cross (×) indicates that the response included some elements that were fabricated or not entirely accurate.	38

List of Acronyms

BVI	Blind and Visually Impaired
LLM	Large Language Models
VLM	Vision-Language Models
AI	Artificial Intelligence
SLIV	Scan-Locate-Interact-Validate
RTAB-Map	Real-Time Appearance-Based Mapping
AR	Augmented Reality
IADL	Instrumental Activity of Daily Living
CV	Computer Vision
RQ	Research Question
ML	Machine Learning
ACM	Association for Computing Machinery
CHI	Conference on Human Factors in Computing Systems
SIGACCESS	Special Interest Group on Accessible Computing
ASSETS	ACM SIGACCESS Conference on Computers and Accessibility
WHO	World Health Organization
UNICEF	United Nations Children’s Fund
TTS	Text-to-Speech
ETA	Electronic Travel Aids
RFID	Radio Frequency Identification
NFC	Near-Field Communication
GPS	Global Positioning System
IMU	Inertial Measurement Unit

YOLO	You Only Look Once
OCR	Optical Character Recognition
NN	Neural Network
MoE	Mixture-of-Experts
CoT	Chain-of-Thought
VL	Vision-Language
GPT	Generative Pre-trained Transformers
VIALM	Visually-Impaired Assistance with Language Models
3D	Three-Dimensional
VSI	Visual Spatial Intelligence
MLLM	Multi-Modal Large Language Model
LM	Language Model
SLAM	Simultaneous Localization and Mapping
FoV	Field-of-View
RL	Reinforcement Learning
UX	User Experience
Redis	REmote DIctionary Server
GPU	Graphics Processing Unit
API	Application Programming Interface
JPEG	Joint Photographic Experts Group
JSON	JavaScript Object Notation
CLIP	Contrastive Language-Image Pre-training
LPU	Language Processing Unit
HTTP	Hypertext Transfer Protocol
REB	Research Ethics Board
VAT	Vision Assistance Technology

Abstract

Grocery shopping is a routine but demanding activity for Blind and Visually Impaired (BVI) individuals, requiring navigation, product identification and verification. While recent large language model (LLM) and vision-language model (VLM)-based tools have improved access to visual information, two key limitations remain: models can produce confident but incorrect descriptions (hallucinations), and most systems rely on turn-based interactions that interrupt continuous, real-world tasks. This thesis explores how assistive artificial intelligence (AI) systems can reduce hallucinations while supporting continuous, in-motion use in environments such as grocery stores. To address this, the thesis introduces ShelfScout, a smartphone-based assistive system that breaks grocery shopping into four stages: Scan, Locate, Interact and Validate (SLIV). Instead of relying on a single model, ShelfScout routes user requests through task-specific pipelines, each designed with its own grounding strategy. The system combines VLMs for object detection and verification with guidance pipelines for aisle-level navigation and hand-level interaction, allowing users to move from initial orientation to product retrieval within a single continuous interaction.

The thesis first presents an initial version of ShelfScout and compares the system-generated responses against three commercial baselines: Be My AI, Meta Ray-Ban smart-glasses and Gemini Live, through a remote user study with eight BVI participants. Results show that ShelfScout produced fewer hallucinated details, while participant feedback highlighted issues such as redundant tasks, unclear spatial phrasing and limited usefulness of certain visual descriptors (e.g., color) for fully blind users. Based on these findings, the system was redesigned to better fit the continuous flow of BVI grocery shopping activities. Key changes include introducing an explicit object presence check, separating far-field navigation from near-field interaction and refining the multi-stage pipeline to improve reliability and clarity. Thus, by structuring how and when models are used, it is possible to improve both reliability and

usability in real-world tasks. The thesis also identifies practical challenges around navigation reliability and response latency that must be addressed before such systems can be deployed and evaluated in fully real-world settings with BVI users.

Abrégé

Les courses alimentaires constituent une activité routinière mais exigeante pour les personnes aveugles et malvoyantes (PAM), nécessitant des capacités de navigation, d'identification et de vérification des produits. Bien que les outils récents basés sur les grands modèles de langage (GML) et les modèles vision-langage (MVL) aient amélioré l'accès à l'information visuelle, deux limitations majeures subsistent : les modèles peuvent produire des descriptions erronées avec un degré élevé de confiance (hallucinations), et la plupart des systèmes reposent sur des interactions séquentielles qui interrompent les tâches continues dans des contextes réels. Cette thèse examine comment les systèmes d'intelligence artificielle (IA) assistive peuvent réduire les hallucinations tout en soutenant une utilisation continue et en mouvement dans des environnements tels que les épiceries. Pour répondre à ces enjeux, la thèse présente ShelfScout, un système d'assistance sur smartphone qui décompose l'activité de magasinage en quatre étapes : Scan, Locate, Interact et Validate (SLIV). Plutôt que de s'appuyer sur un modèle unique, ShelfScout achemine les requêtes des utilisateurs vers des pipelines spécifiques à chaque tâche, chacun étant conçu avec une stratégie d'ancrage propre. Le système combine des MVL pour la détection et la vérification d'objets avec des pipelines de guidage pour la navigation à l'échelle des allées et l'interaction à l'échelle de la main, permettant aux utilisateurs de passer de l'orientation initiale à la récupération du produit dans une interaction continue unique.

La thèse présente d'abord une version initiale de ShelfScout et compare les réponses générées par le système à trois solutions commerciales : Be My AI, les lunettes intelligentes Ray-Ban de Meta et Gemini Live, dans le cadre d'une étude utilisateur à distance menée auprès de huit participants PAM. Les résultats montrent que ShelfScout génère moins de détails hallucinés, tandis que les retours des participants mettent en évidence des problèmes tels que des tâches redondantes, des formulations spatiales peu claires et l'utilité limitée de

certaines descripteurs visuels (par exemple, la couleur) pour les personnes totalement aveugles. Sur la base de ces résultats, le système a été repensé afin de mieux s'aligner sur le déroulement continu des activités de magasinage des personnes PAM. Les principales modifications incluent l'introduction d'une vérification explicite de la présence d'objets, la séparation de la navigation à distance et de l'interaction de proximité, ainsi que l'affinement du pipeline multi-étapes afin d'améliorer la fiabilité et la clarté. Ainsi, en structurant la manière et le moment d'utilisation des modèles, il est possible d'améliorer à la fois la fiabilité et l'utilisabilité dans des tâches réelles. La thèse met également en évidence des défis pratiques liés à la fiabilité de la navigation et à la latence des réponses, qui doivent être résolus avant que de tels systèmes puissent être déployés et évalués dans des contextes entièrement réels avec des utilisateurs PAM.

Acknowledgements

Bringing ShelfScout from a concept to a functional architecture required tremendous amount of support, technical collaboration and real-world feedback.

First, I would like to thank my supervisor, Professor Jeremy Cooperstock, for his ongoing guidance and support throughout this project. While developing ShelfScout, my instinct was usually to throw the heaviest and most complex technologies at the problem to chase higher accuracy. Thank you for consistently challenging that mindset. You taught me how to step back and build from the user’s perspective instead, teaching me that a truly effective system isn’t about integrating the most advanced technology but about finding the best possible way to actually help the people using it. Your advice was invaluable in shaping the direction of this research.

A massive thank you goes to my technical collaborators in the lab: Henrique, Kasra and Adnaan. Building the backend orchestration for ShelfScout would not have been possible without your work on the infrastructure, RTAB-Map and ARKit integrations. Thank you for the debugging sessions, the patience and for helping me figure out how to stitch all of these moving parts together even when the system kept falling apart.

Finally, thank you to my parents, my brother and my friends, for keeping me sane through the countless pipeline iterations, the endless pages of drafting and the numerous technical roadblocks. Thank you for supporting me through it all.

Generative AI Statement

I hereby confirm that I have consulted the [guidelines on the use of generative artificial intelligence](#) issued by McGill's Graduate and Postdoctoral Studies (GPS) and have familiarized myself with my responsibilities as a student regarding the use of generative AI in research. I have consulted with my supervisor and, where appropriate, my department to ensure that any use of generative AI in this thesis complies with both university policies and the norms of my discipline. I attest that all use of generative AI in this thesis adheres to these guidelines. Where necessary, I reference the use of generative AI in my thesis. I understand that any misuse of generative AI may result in disciplinary action under the [Code of Student Conduct](#).

Chapter 1

Introduction

Grocery shopping is a routine activity for most individuals. However, it presents perceptual and physical challenges, combining visual search, product differentiation, spatial navigation and decision-making. Once in a store, shoppers must first orient themselves within the environment, identify required products among the visually similar items and finally, verify features such as flavor, ingredients, price and expiry date. Grocery stores are also visually complex with different aisle layouts and densely packed shelves stocking dozens of products in varied packaging. Products may be occluded by each other, have outdated labels or simply, be in poor condition, all influencing a shopper's purchase decisions on the go. For sighted individuals, each aspect of this task relies heavily on automaticity; they perceive the environment at a glance and take action without experiencing any conscious cognitive strain.

Yet for the Blind and Visually Impaired (BVI), this instrumental activity of daily living (IADL) becomes an “obstacle course”, forcing them to rely on non-visual cues such as tactile feedback to construct mental models of their surroundings and perform tasks such as navigating crowded grocery aisles and circumventing potential hazards along the way. [Khattab et al. \[2015\]](#) highlight that 79% of BVI shoppers find navigating stores difficult and must frequently rely on human assistance to retrieve items from inaccessible shelves, distinguish between specific brands and complete the checkout process. The lack of real-time visual feedback serves as a persistent barrier to their autonomy.

In an effort to better support BVI individuals, recent advances in technology have focused on providing them with more information about their environment, demonstrating a shift from walking canes and guide dogs to computer vision (CV)-based applications to provide detailed visual descriptions about the environment. While promising, these systems depend

on specific hardware setups such as wearable cameras or handheld devices, restricting their practical applicability.

The rise of artificial intelligence (AI) has shifted the approach towards developing accessibility tools. Smartphone-based AI applications are growing to become vital tools for the BVI to access visual information that was previously out of reach. They can now use tools such as [Be My AI](#) (digital visual assistant of [Be My Eyes](#)), [Seeing AI](#) and even general-purpose assistants such as [Siri](#) or [Gemini](#) for scene descriptions, reading labels and assisting them in day-to-day tasks. These systems allow users to issue voice-based queries and receive spoken responses, and can use the device camera to capture images when visual information is required. Their conversational and multi-modal nature has accelerated adoption within the BVI community.

One concern with such technologies is their reliance on large language models (LLMs) and vision-language models (VLMs) to respond to user queries, as these models are prone to *hallucinations*. A hallucination occurs when a model confidently generates false, misleading or fabricated information that is presented as fact but is ungrounded in reality. These errors may sound perfectly logical and plausible, making them difficult to detect without verification. In contexts where the AI system is describing the immediate environment—such as identifying products on a grocery shelf—sighted individuals can cross-check or even challenge AI-generated responses against what they directly observe, catching errors that might otherwise go unnoticed. Blind users lack this independent visual feedback loop and cannot verify environmental information through direct observation. This technical failure point, although a minor inconvenience for the general public, escalates to safety risks for the BVI in complex environments such as grocery stores. Misidentifying allergens or receiving inaccurate walking guidance due to a hallucinated response can lead to unsafe actions and purchases.

Another issue with these AI-assistive systems is their chat-based interaction design. The user asks a question, the system captures and processes an image, and returns an answer, after which the interaction ends. This turn-based structure works well while the user is stationary and can wait for feedback, for instance, when asking to read a document or describe items on a shelf. However, grocery shopping requires continuous guidance across multiple stages: navigating aisles, locating a specific product and grasping it from a shelf. Because the user must remain in motion throughout, the turn-based structure is poorly

suited to support this activity on-the-go. The user must halt their activity, consult the system and resume, breaking the natural continuity of the shopping process.

While each issue is problematic on its own, the combination of hallucinations and the turn-based interaction is limiting in dynamic environments such as grocery stores, where users must both trust the system’s accuracy and rely on it in real-time. Supporting BVI autonomy requires technology that can adapt to user capabilities and situational demands [Herrero-Martín and García-Ballester, 2026]. Although AI holds immense promise for expanding accessibility, prior work by Hu et al. [2026] highlights that assistive agents often produce confident but ungrounded answers when visual evidence is incomplete, a common source of hallucination that points to an issue with how such agents are designed. In addition, Saha et al. [2019] note that their prototype only identifies a target landmark but cannot guide users the remaining distance to it, a gap the authors flag as essential for a complete wayfinding system. Hence, this thesis investigates the following research questions:

RQ1 *How can assistive AI systems be designed to mitigate hallucinations in complex environments such as grocery stores?*

RQ2 *How can these systems support continuous physical mobility in real-time without requiring users to constantly interrupt their activities for re-prompting?*

1.1 ShelfScout

In an attempt to answer the above-mentioned research questions, this thesis proposes “ShelfScout”, a smartphone-based application designed to provide continuous, spoken guidance throughout the grocery shopping process for BVI assistance. At a high level, ShelfScout allows blind shoppers to ask natural questions (“What’s around me?”, “Where are the carrots?”, “Is this the right cereal?”) and receive concise, spatially grounded descriptions that aim to balance real-time question-answering with active navigation support.

Rather than training new machine learning (ML) models from scratch, this work focuses on architectural scaffolding. By structuring off-the-shelf VLMs using specialized prompts and contextual grounding strategies, this thesis demonstrates how system-level engineering can mitigate model hallucinations, thus aiming to address RQ1. Also, to solve the continuous interaction bottleneck, ShelfScout provides users with guidance that supports both

long-range navigation within the store and close-range assistance at the shelf to grasp the object. Users are first guided toward the correct aisle to reach their target product. Once at the correct location, the system provides real-time audio feedback to support locating and grasping the object, thus aiming to address RQ2.

1.2 Thesis Structure

The organizational structure of this thesis is as follows, with a short overview of each chapter:

Chapter 2—Literature Review provides a comprehensive overview of the background material and related work in AI-assisted accessibility tools, establishing the context for our approach.

Chapter 3—System Design introduces ShelfScout, defining its primary design goals, interaction model and high-level architecture, while clarifying the specific scope of this thesis and its contributions within the broader collaborative project.

Chapter 4—Preliminary User Study outlines the initial system design and details a remote qualitative study conducted with BVI users. This study evaluates how participants perceive and respond to simulated AI-generated audio feedback, inspiring foundational design requirements for ShelfScout.

Chapter 5—System Evolution describes how the system architecture evolved in response to technological advances and user feedback from the preliminary study. It details the final pipeline implementation and presents the evaluation of ShelfScout by researchers and its limitations.

Chapter 6—Conclusion summarizes the main contributions of this thesis and discusses the main results. It identifies potential future research directions for improving the methodologies developed in this thesis.

1.3 Contributions of Collaborators

ShelfScout was built as a collaborative project and the responsibilities of each team member are outlined below to clarify the specific scope of this thesis:

- **Mansi Dhanania (Author of this thesis):** designed and implemented the vision pipeline, including VLM integration, object detection and response filtering to ground outputs, described in Chapter 3. Served as a co-investigator for designing and conducting the preliminary user study described in Chapter 4, including study design, participant recruitment, data collection and analysis of quantitative and qualitative feedback. Designed the orchestration layer that routes information through continuous feedback loops between the vision and guidance pipelines, highlighted in Chapter 5.
- **Henrique J. P. Oliveira (PhD Candidate):** co-designed the initial system architecture and infrastructure (Docker, Redis, n8n) described in Chapter 3, developed user personas and led HCI evaluation. He served as a co-investigator for the user study described in Chapter 4. He also provided extensive feedback on the overall thesis.
- **Kasra Jabbari (Master’s):** designed and developed the far-field navigation pipeline, described in Chapter 5, that enables BVI users to traverse the store environment autonomously, handling localization and path planning in indoor spaces.
- **Mohammad Adnaan (Master’s):** implemented the ARKit near-field guidance pipeline described in Chapter 5, providing hand-level guidance for users to grasp objects.
- **Professor Jeremy Cooperstock (Supervisor):** provided guidance on system design, research direction and project oversight throughout the development of ShelfScout. He also provided feedback on drafting the overall thesis.

The author acknowledges that ShelfScout was developed in collaboration with the above-mentioned team, and its high-level design and deployment were collective efforts. While this thesis presents the overall architecture of ShelfScout, the primary focus is on the design and orchestration of the multi-modal pipeline layers. Technical details regarding the front-end mobile and web application development, as well as the development of the far-field navigation and near-field AR-based guidance sub-components, are beyond the scope of this thesis and are the responsibility of the respective team members.

Chapter 2

Literature Review

Assistive technologies for the BVI are changing from static, single-task tools to interactive assistants capable of autonomous reasoning and complex task execution. [Mack et al. \[2021\]](#) surveyed research papers in ACM CHI and ASSETS conferences and reported that accessibility research has grown to make up almost 8% of all published papers at CHI as of 2019. Analyzing accessibility-focused papers from both venues between 2010 and 2019, they found that 43.5% targeted the needs of BVI users. Despite this growth in research, sustained adoption of assistive technologies among BVI individuals remains a challenge. [Barak Ventura et al. \[2025\]](#) report that assistive technology abandonment has ranged from 17% to 50% for optical devices used by people with blindness and low vision, attributing this pattern to design choices that overlook user needs, unreliable manufacturer support and insufficient guidance for users to learn the tools effectively. This points to a persistent gap between technological capability and real-world adoption among BVI individuals.

In order to understand the factors contributing to this gap, this literature review first evaluates the current state of BVI assistive research. It examines the evolution of BVI aids and highlights the unmet socio-technical needs of BVI individuals, particularly during grocery shopping. To motivate ShelfScout’s design decisions, the literature review will then analyze the promise and limitations of LLMs and VLMs in accessibility, focusing heavily on model reliability and the friction caused by conventional turn-based interaction designs. By identifying these systemic bottlenecks, this chapter establishes the structural requirements for the orchestration and navigation frameworks proposed via ShelfScout in this thesis.

2.1 BVI Accessibility

Assistive technologies for the BVI community have progressed from human-dependent support to autonomous, AI-driven tools. This trajectory has been driven by a continuous effort to enhance user independence [Du and Bulusu, 2021]. An assistive tool’s success depends not only on its computational capability, but also on its ability to function effectively beyond controlled lab environments, in unpredictable, real-world settings. In addition, human factors such as social acceptability, financial cost and cognitive load of using the tool, all influence whether an aid is actually adopted or abandoned by users [McDonnall and Sessler Trinkowsky, 2025]. The following section traces how these factors have shaped the evolution of BVI assistive technology.

2.1.1 Evolution of Assistive Tools for BVI

Early assistive technologies focused on substituting visual information with auditory or tactile feedback. For physical mobility, tools such as the white cane and guide dogs were used to detect obstacles in an individual’s path. Due to their simplicity and reliability, they remain a primary tool for BVI orientation and mobility. For accessing written content, the Braille system allowed individuals to read and write via tactile patterns. Over time, with a shift toward digital screens, screen readers became the cornerstone of digital accessibility, allowing users to navigate web layouts via synthesized text-to-speech (TTS). Although effective for text-based interfaces, these tools provide no feedback regarding the user’s physical environment or help them interact with their surroundings in real-time [Naayini et al., 2025].

To bridge this information gap, electronic travel aids (ETAs) and camera-based systems emerged. Initial efforts relied on augmenting physical infrastructure such as radio frequency identification (RFID) tags, near-field communication (NFC) stickers and Bluetooth beacons to “pin” objects and locations within mapped, controlled environments. However, these systems translated poorly to real-world deployment because they required modifying existing physical spaces and maintaining constantly updated digital representations of the environment—an impractical burden in real-world settings. Thus, academic research also explored user-wearable hardware, resulting in prototypes such as specialized bar code scanners and early “smart canes”. These tools also suffered from poor usability due to their bulky

form factors and high manufacturing costs. They required precise, manual positioning of bar code labels, making them lack the flexibility needed to navigate spontaneous daily activities [Lavric et al., 2024].

A paradigm shift occurred with the ubiquity of the smartphone. By leveraging built-in cameras, screen-readers (such as [iOS VoiceOver](#) and [Android TalkBack](#)) [Rodrigues et al., 2020] and internal sensors including global positioning systems (GPS) for outdoor positioning and inertial measurement units (IMUs) for user orientation, smartphones enabled diverse software solutions. Mobile applications such as [Microsoft’s Seeing AI](#) and [Google’s Lookout](#) could run CV algorithms and convey information via synthesized speech [Patel et al., 2024, Azam and Khan, 2025]. Local object detection frameworks such as [You Only Look Once \(YOLO\)](#) [Redmon et al., 2016] were used to identify daily items or currency, alongside optical character recognition (OCR) engines such as [EasyOCR](#) and [Tesseract](#) to extract text [Smith, 2007, Marquez-Carpintero et al., 2026]. The burden of environmental awareness shifted from physical infrastructure to software applications.

Despite their utility, these CV-based applications suffered from severe architectural limitations. They relied on isolated functionalities, requiring users to manually switch between predefined modes (e.g., switching from an “OCR mode” to an “object recognition mode”), increasing cognitive load. Fernando et al. [2025] highlighted that such tools offered limited support for complex, multi-step tasks that involved planning, navigation and verification. BVI participants also reported concerns about verbosity, irrelevant details and lack of control over the level and type of information provided. As a result, users could receive information that is factually correct but difficult to translate into real-world action. Also, these applications were constrained by small training datasets and rule-based logic, making them incapable of conveying high-fidelity information about visual scenes [Gurari et al., 2018].

Now, the rise of AI has redefined accessibility. For the BVI community, smartphone-based AI applications have become vital tools for accessing information that was previously out of reach. Modern AI assistants are interactive, conversational and multi-modal [Fernando et al., 2025]. This change from simple rule-based recognition to complex reasoning [Wei et al., 2022] is driven by the development of the “Transformer” architecture in neural networks (NN) [Vaswani et al., 2017]. Unlike the CV systems, LLMs and VLMs are trained on a massive corpus of text and images from the internet, which allows them to generate natural language responses through “next-token prediction”. Techniques such as mixture of experts

(MoE) allow these models to trigger specific sub-networks for different tasks and chain of thought (CoT) prompting lets the model break down complex visual queries into logical steps before providing an answer [Shazeer et al., 2017, Wei et al., 2022]. This results in a system that responds almost indistinguishably from a human and, is becoming commercialized into mainstream accessibility applications such as [Be My AI](#) (digital visual assistant of [Be My Eyes](#)), [Seeing AI](#) and, even general-purpose assistants such as [Siri](#) or [Gemini](#). Users can issue spoken queries, capture images and receive synthesized audio feedback in real-time.

The accessibility and portability of smartphones has also accelerated adoption, making AI-based vision assistance one of the most commonly used forms of assistive technology among BVI users [Naayini et al., 2025]. For many, these tools have become integrated into everyday tasks such as shopping, cooking and traveling. [Fernando et al. \[2025\]](#) report high smartphone adoption among BVI individuals, who are quickly relying on a single, multi-functional device for communication, navigation, text access and environmental awareness. Users express a preference for discrete, portable systems that fit into an existing routine rather than requiring additional devices [Shinohara and Wobbrock, 2011].

New form factors are also emerging. Hands-free smart-glasses offer an alternative to handheld smartphones, enabling users to access visual assistance without occupying their hands. Devices such as the [Meta Ray-Ban](#) smart-glasses or the [Envision](#) glasses utilize the same underlying LLMs and VLMs but shift the camera—required to capture visual data—to the stems of glasses, creating wearable technology that provides a more natural and egocentric perspective of the world [Xiao et al., 2026]. While promising, this technology is still quite expensive and poses hardware limitations such as poor battery-life to run state-of-the-art multi-modal models locally [Naayini et al., 2025, Lee et al., 2025].

2.1.2 Prior Work in BVI Shopping Assistance

Looking at shopping assistance, a range of systems have attempted to address different parts of the grocery shopping activity. Prior work such as [Trinetra \[Lanigan et al., 2006\]](#) and [ThirdEye \[Carroll et al., 2017\]](#) combined bar code scanning and RFID tags with mobile interfaces to help BVI users identify products in-store. These systems showed the value of mobile-based aid for shopping but depended on specialized tags and controlled environments that limit generalizability to all stores. [Liao et al. \[2024\]](#) developed [EasyGO](#), a system focused on indoor navigation to provide visually impaired individuals guidance through grocery

environments based on store maps and predefined points of interest. To support product acquisition, EasyGO displayed directional arrow cues on a smartphone screen once a user was near the target shelf, guiding them toward the item’s location. However, a fully blind shopper cannot see an on-screen arrow and would need to rely on the accompanying audio cues alone, which the study found participants used only as a secondary channel compared to the screen. Also, EasyGO provided no mechanism to confirm that the item reached was in fact the intended product. This leaves a fully blind shopper unable to independently locate the correct item or confirm its identity once acquired, without relying on sighted assistance.

The emergence of vision-language (VL) assistants (e.g., GPT-4V-based applications such as Be My AI [Be My Eyes, 2023]) offered new perspectives for shopping support. Such systems can handle free-form questions about scenes, products and labels, and they operate on smartphones, aligning with BVI preferences for single-device, conversational tools [Fernando et al., 2025]. Applications such as SeeingGrocer [Tang et al., 2025] advanced by integrating sensing hardware such as an RFID wristband and a smart shopping cart, with a conversational LLM-based assistant. Unlike its predecessors, SeeingGrocer moved beyond object detection to offer context-aware environment understanding. It can interpret complex visual scenes to describe store aisles and retrieve specific product features such as pricing or nutritional data. This multi-modal approach also provides stepwise guidance that mirrors the experience of shopping with a sighted companion. However, the system still faced several deployment barriers. Its reliance on ArUco markers for indoor localization and RFID-tagged shelves required infrastructure accommodations that are not yet a standard in most retail stores. Also, while the conversational AI is flexible, the system’s performance remains dependent on the accuracy of the underlying VLM, which can still struggle with fine-grained text recognition and be prone to hallucinations, leading to misinformation. Recent benchmarks such as VIALM [Zhao et al., 2024b] reveal that even state-of-the-art models fail to provide well-grounded and detailed guidance, giving incorrect responses in cluttered environments [Liu et al., 2024]. These issues are problematic in grocery shopping, where even small inaccuracies can lead to wrong purchases and reduce trust in the system.

2.1.3 Unmet Needs in BVI Grocery Shopping

Gamage et al. [2023] conducted a large-scale review of 646 research papers on assistive technologies for visually impaired users and compared them against user-reported function

priorities. Their analysis showed only a weak positive correlation between research activity and what BVI participants actually identified as high-impact need. Text recognition, scene understanding and obstacle detection ranked among the most valued capabilities. However, most published research focused on narrow CV benchmarks and constrained lab scenarios rather than real-world deployment. Also, navigating complex retail environments and locating specific products on shelves were identified as underexplored despite their importance in participants’ daily lives. [Lee et al. \[2021\]](#) highlight how grocery shopping exemplifies this disconnect: while users identify shopping and product verification as important domains, few systems provide end-to-end support — from initial orientation to product acquisition and validation. Existing systems target a single component of the shopping workflow, for example, aisle-level navigation or label reading or object detection, without providing a continuous experience from initial orientation to final product validation.

In addition, current AI solutions such as those found in Meta Ray-Ban or general-purpose VLMs provide generic scene descriptions (e.g., “You are in a cereal aisle”) that fail to provide the granular data necessary for a purchase. [Lee et al. \[2021\]](#) showed that BVI shoppers rely heavily on details such as price, freshness and expiration dates, and that the generic image captions were insufficient for making purchase decisions. When a VLM provides a “plausible” response or worse, fabricates a detail about an allergen or a price, the lack of a visual feedback loop for the BVI user transforms a minor technical error into a significant safety risk or a loss of trust. [Fernando et al. \[2025\]](#) reported that product verification is a constant challenge, forcing BVI shoppers to rely on store staff or sighted companions, increasing cognitive load during decision-making under time constraints. These studies highlight that successful shopping support goes beyond high-level scene descriptions. It requires that users be provided with precise, understandable information about specific items that can influence their purchase decisions.

[Agrawal and Hayes \[2026\]](#) highlight how the physical complexity of the store environment also remains an unsolved hurdle for existing software. Grocery stores are visually complex environments with densely packed shelves, dynamic layouts and thousands of products varying in shapes and packaging, which creates edge cases that challenge even the most advanced AI architectures. Produce aisles test shape-based recognition of unlabeled items such as carrots, while cereal aisles demand precise OCR for branded, text-heavy boxes. Products may also be partially occluded, rearranged daily or visually similar (e.g., differentiating boxes of

oats from cornflakes). Beyond object identification, grocery shopping also requires three-dimensional (3D) spatial navigation and physical interaction. Users need to move through aisles, orient themselves relative to shelves and reach for items. Because grocery shopping is not a static task but a continuous interaction between the user, the AI and the physical shelf, current “chat-based” AI paradigms usually fail. They lack the spatial grounding and real-time guidance necessary to help a user move from “identifying” a product to “locating and grasping” it [Ruan et al., 2026].

To address the last-meter guidance problem, an assistive system needs to understand how objects are laid out in space. Yang et al. [2025] attempted to solve this through VSI-Bench, a benchmark designed to evaluate “visual-spatial intelligence” in multi-modal LLMs (MLLMs), testing how well MLLMs see and remember physical environments. They found that traditional ways of helping LLMs think, such as asking them to explain their reasoning step-by-step, do not usually work for spatial tasks and can actually make the models more confused. They highlighted that the AI performs better when it is forced to create a cognitive map or a mental picture of the physical space. For BVI assistance, this means an AI should not just describe what is directly in front of the camera; instead it needs a consistent bird’s-eye view and understanding of the store that stays the same even as the user turns around. However, these mental maps were tested in simplified, digital simulations rather than the unpredictable environment of a real-life grocery store. Also, the study focused on how models recall space based on static images, leaving a gap in understanding how an AI would maintain this map in real-time as a user moves through a crowded aisle with changing lighting and moving people. These challenges make grocery shopping an ideal testbed for studying what AI-based assistance must do beyond “describe the scene” to actually support BVI autonomy.

2.2 Language Models and Hallucinations

Liu et al. [2024] and Dang et al. [2025] show that one of the biggest challenges in AI responses is hallucination. In assistive tools, hallucinations are very problematic because users depend on the system’s output to make decisions in the physical world. For blind users, hallucinations can lead to them attempting to grasp objects that are not present, misidentifying products or following navigation instructions that do not correspond to the real-world environment captured through the camera. These inaccuracies are not always perceivable

as hallucinated responses are delivered in the same confident tone as correct ones, with convincing details. This risk is compounded in multi-modal systems where a hallucination in one modality can propagate through the entire pipeline. For example, an incorrect visual interpretation can lead to errors in textual descriptions, which are then converted into audio feedback. This compounding effect can aggravate the impact of a single misinterpretation, creating a completely false mental model of the environment for the user.

Liu et al. [2024]’s survey characterized several reoccurring failures: object fabrication,¹ feature hallucination,² spatial hallucination³ and logical hallucination.⁴ These hallucinations are a by-product of how generative models are designed, arising from the training objective, which optimizes next-token probabilities conditional to both text and image embeddings, and prioritizing fluency, indicating how natural the sentence sounds, over perceptual fidelity, indicating how accurate the sentence is to the image. As a result, LMs may confidently produce detailed answers even when the visual evidence is ambiguous or missing [Dang et al., 2025]. Integrating visual data with language generation still demonstrates VLMs leaning towards internal linguistic patterns over grounded evidence. A model’s “reliability”, i.e., its ability to provide consistent answers, also does not guarantee accuracy. It can be confidently and consistently wrong.

GPT-4V-based systems such as Be My AI [Be My Eyes, 2023] highlight cases where the system confidently misidentifies objects or misreads text. In grocery environments, Tang et al. [2025] document hallucinated brands, incorrect shelf positions and made-up product variants when aisles are densely stocked or items are partially occluded. Users described feeling misled when systems provided long, confident explanations that later turned out to be inaccurate, eroding trust even when subsequent responses were correct [Nikanfar et al., 2025]. These findings echo concerns that hallucinations in VLMs are a safety and usability issue when outputs are used to guide physical actions rather than answer abstract questions [Liu et al., 2024].

These limitations motivate modifications to constrain or steer VLM behavior, including post-hoc verification and prompt design [Dang et al., 2025]. Muennighoff et al. [2025] introduce the “s1 model”, which utilizes a technique called “budget forcing” to improve reasoning

¹Object fabrication refers to the model describing items that are not present in the image.

²Feature hallucination can include describing incorrect colors, brands or prices.

³Spatial hallucination refers to the model describing incorrect object locations.

⁴Logical hallucination refers to the model drawing incorrect conclusions from partial evidence.

reliability. s1 is trained on high-quality reasoning text and uses extra computation at test-time to “deliberate”. By forcefully extending the model’s thinking process through “wait” prompts, the system is encouraged to double-check its initial logic and self-correct errors before providing a final answer. This approach demonstrates a “test-time scaling” effect where accuracy correlates with the amount of “thought” allowed, providing a technical blueprint for assistive tools that are more dependable and less prone to the high-stakes mistakes often found in cluttered retail environments.

Prompting techniques such as explicitly instructing the model to avoid guessing, to state uncertainty or to focus only on verifiable evidence can also reduce some hallucinations in controlled benchmarks. Prompt engineering is a widely adopted technique for shaping the behavior of LLMs and VLMs without modifying their parameters [Schmidtová et al., 2025]. By carefully crafting instructions in prompts, providing examples or specifying output formats, LLMs and VLMs can be encouraged to exhibit desirable behaviors such as step-by-step reasoning, uncertainty expression or grounded description [Dang et al., 2025]. For instance, CoT prompting, which asks models to verbalize intermediate reasoning steps, has been shown to improve performance on reasoning tasks and can reduce certain types of hallucination by approximately 10-20% in benchmark evaluations, demonstrated by Dang et al. [2025]. Similarly, prompting VLMs to “only describe what is clearly visible” or to output structured JSON with explicit uncertainty fields can help constrain visual hallucinations in simple scenes [Liu et al., 2024].

Despite these benefits, Schmidtová et al. [2025] and Serge Liatko [2025] highlight fundamental limitations of prompt-based control in complex multimodal and multi-task settings. First, VLMs often respond better to concise, focused prompts than to long, exhaustive instructions. As prompts grow to encode all possible edge cases, models may ignore parts of the instruction or revert to generic patterns learned during pre-training. Serge Liatko [2025] suggested diminishing returns or even performance degradation when prompts exceed certain lengths, as models struggle to maintain attention over all constraints. Longer prompts intended to enforce grounding and caution can produce more speculative, verbose answers rather than fewer hallucinations [Schmidtová et al., 2025].

Prompt engineering also scales poorly across varying tasks and environments. A prompt that works well for high-level scene description may be inappropriate for near-field reaching instructions, where concise, egocentric commands are needed, or for label reading, where

structured text extraction is more important than narrative detail [Lee et al., 2021]. As summarized by Dang et al. [2025], attempting to maintain a single, monolithic prompt that handles all intents risks conflating these distinct interactions, making it difficult to reason about when and why hallucinations occur. Also, models do not possess true symbolic reasoning; they cannot reliably enforce global rules such as “never invent objects” across all responses, especially when visually ambiguous input conflicts with training parameters [Liu et al., 2024].

Finally, Serge Liatko [2025] demonstrate how prompt-based behavior is fragile with respect to conversational drift. Small changes in user phrasing, image composition or conversation history can lead to different interpretations of the same instruction, resulting in inconsistency in responses sticking to safety prompts. In assistive use, where users may ask follow-up questions, rephrase requests or move through the store while interacting with the system, relying on prompts to prevent hallucinations is risky. These limitations suggest that prompt engineering should be treated as one layer of a broader safety strategy rather than as the primary mechanism for controlling VLM behavior.

Liu et al. [2024]’s survey consistently found that such measures offer only partial mitigation and that VLMs still hallucinate, especially in cluttered or unfamiliar domains. Hallucinations in assistive AI should be understood not only as a model-level issue but as a system design problem. The way tasks are framed, decomposed and routed through different components can influence whether a system guesses, defers or grounds its responses in perception. In contexts where confidence of an AI-generated response is crucial for BVI users, designing AI systems that know when not to answer or convey their uncertainty can be as important as improving raw model performance.

2.3 BVI Guidance

Effective navigation remains one of the biggest challenges for BVI individuals, requiring both global orientation⁵ and local guidance [Bamdad et al., 2024].⁶ Traditional assistive navigation such as EasyGo by Liao et al. [2024] has relied on GPS for outdoor environments and indoor solutions such as Wi-Fi finger-printing, Bluetooth beacons or pre-built store

⁵Global orientation refers to identifying one’s position in an environment.

⁶Local guidance refers to instructions to help traverse through the environment.

maps. However, grocery stores present unique difficulties: narrow aisles, frequent restocking, temporary product placements and dynamic obstacles (e.g., shoppers, carts) that make static maps unreliable, while GPS completely fails in indoor environments due to signal attenuation through building materials and interference caused by reflective surfaces such as metal shelves [Hailu et al., 2025]. Hence, recent research has shifted away from external infrastructure toward user-centered spatial sensing.

Bamdad et al. [2024], Xu et al. [2025] and several other recent research have increasingly explored visual simultaneous localization and mapping (SLAM) techniques adapted for BVI navigation. RTAB-Map [Labbé and Michaud, 2019] and similar visual-SLAM systems generate real-time 3D point cloud maps from egocentric camera feeds while simultaneously tracking user position, enabling obstacle avoidance and path planning without pre-existing infrastructure. Bamdad et al. [2024] surveyed SLAM-for-BVI applications and highlighted their potential for the “last-meter” guidance, i.e., guiding the user to a distance of 1-2 meters from the target location, but noted several limitations for grocery contexts. First, most systems focused on safe path-finding around obstacles rather than precise object localization on shelves [Xu et al., 2025]. SLAM outputs are usually low-level 3D coordinates and pose estimates [Labbé and Michaud, 2019] that require translation into natural language instructions such as “turn left, walk 2 meters, reach at waist height” [Lin et al., 2026]. Also, [Bamdad et al., 2024] pointed out that phone cameras are ill-suited for the “last-meter” task as they suffer from motion blur and limited field-of-view (FoV) in tight spaces. In tight grocery aisles, these factors prevent the system from capturing enough environmental landmarks to maintain a stable map, leading to frequent tracking failures.

This divide between geometric tracking and semantic understanding splits current navigation assistance into two distinct paradigms: egocentric and exocentric [Gamage et al., 2024]. An egocentric point of view refers to a first-person view through either a phone or smart-glasses, while an exocentric view-point refers to a third-person view via overhead maps. Egocentric approaches align naturally with BVI interaction patterns, which includes the user holding their phone, speaking an instruction to the system and receiving responses, but struggle beyond arm’s reach—most VLMs provide vague spatial descriptions (“Carrots are on the top right shelf.”) without metric distances and sequential instructions [Liu et al., 2024]. Exocentric solutions such as NaviLens tags⁷ [NaviLens, 2022] offer precise aisle-level

⁷NaviLens tags are high-contrast, camera-readable, color matrix bar codes attached to signage that can

routing via real-time audio guidance, but require infrastructure and break down at product-level. Because each tag must be physically affixed to a fixed location (e.g., a product or a shelf), this approach works well for guiding a user to the general area near the required product, but does not scale to labeling every individual product on the shelf, leaving the “last-meter” identification and grasping problem unsolved. Deep reinforcement learning (RL) approaches such as work by [Lu et al. \[2021\]](#), have explored end-to-end policies from raw video to voice commands but generalization across store layouts remains poor.

Hence, grocery aisles expose a specific gap in both egocentric and exocentric approaches as they target the “mid-range” problem of guiding users across shorter distances of 2-10 meters to a product zone and then transitioning to fine-grained reaching. Some prior work does attempt to address this transition. [Visually Impaired Robot-Assisted Shopping \(VIRAS\)](#) is a motorized shopping cart that uses onboard stereo cameras to build a 3D map of the store, physically navigate the user to the correct aisle, and then track the user’s hand to guide them to the target product on the shelf. However, it relies on specialized carts that a store would need to purchase and maintain, making it impractical for everyday, independent BVI shopping. [Hao et al. \[2022\]](#) propose a wearable, vision-only system that localizes a target object in 3D space and continuously re-estimates the user’s trajectory toward it as they walk, correcting for drift when the user steps around obstacles. This again requires an additional piece of wearable hardware, and its object detection module relies on YOLO, which is trained on a fixed set of object labels and does not generalize to the diversity of grocery products a user might request. The shift towards open-vocabulary VLMs creates a different problem: current models hallucinate distances and shelf positions in cluttered shelves [[Zhao et al., 2024b](#)] and SLAM systems localize but generate coordinate lists that are not directly usable to generate guidance instructions [[Labbé and Michaud, 2019](#)]. Approaches combining VLMs with SLAM remain underexplored without specialized sensors [[Gamage et al., 2024](#)].

ShelfScout addresses these gaps by integrating egocentric VLM object detection for near-field guidance with RTAB-Map localization for in-store aisle traversal. SLAM pose sequences are transformed into prioritized “walk/turn” sequences, while routing between workflows reduces hallucination when objects are out-of-frame. The following chapter presents the system architecture and explains how decomposing tasks helps coordinate perception and mapping to support autonomous grocery shopping.

be read from over 30 meters away and even at angles up to 160 degrees.

Chapter 3

Preliminary System Design

This chapter introduces ShelfScout, the assistive AI system developed and evaluated in this thesis. Chapter 1 posed two questions: RQ1 asked how assistive AI systems can be designed to mitigate hallucinations or convey uncertainty in complex environments to ensure user safety and, RQ2 asked how these systems can support continuous physical mobility without forcing users to repeatedly interrupt their activity to consult the system.

Looking at prior work to guide the design of our system to address these questions, we highlight two limitations in existing systems. The first is hallucinations, elaborated upon in Section 2.2 as the model’s tendency to generate plausible but incorrect responses when visual information is ambiguous or incomplete. Second, Section 2.3 showed that egocentric VLM-based approaches and exocentric mapping approaches for guidance do not enable the continuous guidance loop that physically walking through a store and reaching for a product actually requires [Gamage et al., 2024].

ShelfScout responds to both observations with a single design decision: rather than treating grocery shopping as one interaction between the user and AI, the system decomposes it into a small number of task-specific stages, each with its own grounding strategy and its own interaction flow. This chapter describes the system’s design goals, interaction flow and high-level architecture, providing context for the detailed pipeline and architectural decisions discussed following that.

3.1 Design Goals

ShelfScout is organized around three design goals:

1. **Grounded responses** (*addresses RQ1*): Rather than depending on a single VLM to handle all visual queries, ShelfScout implements explicit intent-based checks and fallbacks to prevent hallucinated descriptions or guidance [Liu et al., 2024]. When reliable visual grounding is not available, the system defers rather than forcefully generating a suitable response to the query, addressing the overconfidence problem in VLMs [Be My Eyes, 2023, Zhao et al., 2024b].
2. **Sequential guidance** (*addresses RQ2*): Grocery shopping requires multiple sequential actions: environment scan, orientation, object localization, navigation and verification [Gamage et al., 2023, Lee et al., 2021], that do not fit into a single question-answer exchange. ShelfScout decomposes these into separate tasks with their own pipelines, allowing the system to keep guiding the user across a transition (e.g., from “which aisle” to “which shelf” to “grasp now”) without requiring the user to re-initiate a new interaction at each step [Bamdad et al., 2024].
3. **Modular architecture** (*addresses RQ1&2*): Recognizing rapid LLM/VLM developments, ShelfScout uses open-source LMs orchestrated through containerized microservices [Labbé and Michaud, 2019]. This modular design allows easy substitution of VLMs without disrupting the workflow or user experience.

3.2 SLIV Framework

ShelfScout decomposes the grocery shopping activity into four distinct tasks: Scan (scan the environment to orient the user), Locate (identify objects of interest in the immediate surroundings), Interact (guide the user to reach the desired product location) and Validate (confirm whether the desired product was reached and grasped), referred to as the “SLIV” framework throughout this thesis. The system’s architecture mirrors this decomposition: each of the four tasks is implemented as a separate pipeline, with its own model configuration and interaction style.

This decision was influenced by the results from several prior research studies in this area. Kulyukin and Kutiyawala [2010] argue that finding a product in a store breaks down into three distinct “performance gaps”: store navigation, product search and product identification, because each is enacted differently by the shopper and, hence, needs to

be processed differently by any assistive system. [Deryagina et al. \[2023\]](#) validate this from the user’s side: in walking-tour interviews, BVI shoppers described locating a basket as an unstructured “random search” distinct from the targeted search for a specific product, and described label-reading near crowded shelves as a separate, overwhelming task in its own right. SLIV extends [Kulyukin and Kutiyawala’s](#) three-way split into four stages. Since [Balconi et al. \[2024\]](#) show that BVI users face difficulty in constructing mental spatial maps of stores, leading to disorientation, isolating *Scan* from *Interact* gives ShelfScout the ability to manage that disorientation by awaiting the user’s confirmation to initiate continuous guidance. Also, while BVI shoppers rely heavily on tactile exploration for product identification, *Validate* is designed to augment that physical interaction with fine-grained details such as price, expiry date or fine print. ShelfScout parses this information from real-time images captured through a smartphone and conveys the same to BVI users through voice responses. In terms of system architecture, the task-level separation gives each task a different grounding requirement rather than one shared by all four. The literature review identified hallucinated fine-grained details as a recurring failure point. For example, [Tang et al. \[2025\]](#) reported hallucinated items or shelf locations as the model reasons over a wide, cluttered field of view, and [Lee et al. \[2021\]](#) found that generic scene captions are insufficient for the fine-grained details a purchase decision actually requires [[Liu et al., 2024](#)]. Treating *Validate* as its own stage, with its own grounding requirement, is SLIV’s response to that specific gap.

Table 3.1 summarizes each task, its intent and its design rationale.

3.3 Hardware Form-Factor

The choice of a smartphone-based modality over dedicated wearable hardware (e.g., smart-glasses or head-mounted cameras) is driven by three factors identified in assistive technology literature: ubiquity, social acceptability and computational versatility [[Wilson et al., 2022](#)].

Smartphones have transitioned from general-purpose devices to “indispensable assistive tools” that BVI users already own and carry, eliminating the “abandonment risk” associated with specialized hardware [[Senjam, 2021](#)]. Also, research into the social overhead of assistive tech shows that BVI users prefer mainstream devices that do not accentuate being visually impaired in public or carry the social stigma of medical-looking wearables [[Senjam, 2021](#),

Table 3.1: SLIV Task Decomposition

Task	Intent	Query Examples	Rationale
Scan	Description of the environment	“What’s around me?”, “Describe the scene”	Orient the user in their immediate surroundings
Locate	Detect objects	“Do you see apples?”, “Where’s the milk?”	Identify the presence and location of target objects in the user’s vicinity
Interact	Navigation and Reaching guidance	“Guide me to the apples”, “How do I reach it?”	Sequential guidance to the required aisle and assistance in reaching the object on the shelf
Validate	Product details such as price, expiry date, quantity etc.	“Is this the right one?”, “What’s the price?”	Fine-grained verification for decision-making

[Shinohara and Wobbrock, 2011]. In addition, smartphones offer a multi-modal sensor suite, including cameras, IMUs and dedicated AI chips that can match the performance of stand-alone vision systems while allowing for software updates as VLMs evolve [Park, 2011]. By using the smartphone as the primary hardware interface, ShelfScout ensures that the system is not only technically capable but also economically and socially accessible for daily use in crowded retail environments.

3.4 User Experience

The ShelfScout User Experience (UX) is designed as a conversational loop that mirrors the natural cognitive process of a shopper. The system allows users to move fluidly between the stages of the SLIV framework through natural language. The user first enters a store and triggers the system via a double tap [Kane et al., 2008, Alonso et al., 2008]. The system starts voice recording the user’s verbal query with automatic “end-of-speech” detection. Once the user provides a verbal query (e.g., “What is around me?”), the system captures the current camera frame, ensuring the visual data is synchronized with the user’s question [Bigham et al., 2010, Kane et al., 2011]. The user waits for a brief processing interval while the system routes the query through the SLIV pipeline. ShelfScout then provides an audio

response via synthesized speech (e.g., “You are in the produce section of a grocery store. Are you looking for something specific?”).

Based on the feedback, the user may prompt the system again. If they wish to look for a specific object (e.g., oranges, apples, cereal, pizza), they can ask the system to guide them to that object. Depending on whether the object of interest is visible in the camera’s view, the system will automatically route the request along the respective processing pipelines for near-field or far-field guidance, which both work in continuous loops to ensure the user receives real-time feedback until the task has been completed [Kim, 2022]. They may also re-prompt the system (e.g., “Am I close to the apples yet?” or “Is my hand near the cereal boxes?”) according to their comfort. In response, the system will provide updated spatial feedback about the target object’s location relative to the user’s current position (e.g., “Yes, the apples are to your right, within arm’s reach.”). This allows the user to use the system as a real-time spatial “proxy”. Once an item is acquired, the user may use the validate phase. By holding the product in front of the camera and asking for product specifics (e.g., “Give me some details about this product?” or “Is this Kellogg’s Rice Krispies?”), the user gains access to fine-grained information that is often inaccessible through tactile exploration alone. By allowing the user to re-prompt at any moment, ShelfScout functions as a collaborative agent that updates its responses based on the latest available visual inputs and user queries. This interaction model reflects a broader pattern identified in prior work: Gamage et al. [2023] noted that BVI individuals preferred natural language question-answering and follow-up queries over rigid menu structures, especially for contextual scene understanding and task-oriented assistance. The interaction affordances of modern LLMs and VLMs closely align with this finding as they can process images and text to generate context-aware responses.

3.5 Architecture

ShelfScout’s preliminary architecture is designed to support the SLIV framework by routing queries through task-specific processing pathways based on intent classification. The backend is deployed as a suite of Docker containers orchestrated through a self-hosted docker instance of n8n, an open-source workflow automation tool that manages inter-service communication and state persistence. n8n acts as the orchestration tool to manage the SLIV workflow by routing transcripts and images sent from the front-end through intent classification and

task-specific visual reasoning pipelines. This layer coordinates Application Programming Interface (API) calls to external services such as [Groq](#) for LLM/VLM inference and internal services such as docker containers running the guidance pipelines.

All stateful data such as user queries, session context and intermediate outputs are stored in a locally running instance of [REmote DIctionary Server \(Redis\)](#), allowing conversational continuity across multiple interactions. By containerizing each component, ShelfScout ensures reproducibility and deployment across differing hardware (local GPUs, cloud APIs).

3.5.1 Intent-Based Classification

When the user completes their button press interaction (press-speak), the front-end sends the transcribed text along with the Joint Photographic Experts Group (JPEG)-encoded camera frame image to n8n via a web-hook endpoint.

To route queries into the appropriate SLIV pathways, ShelfScout employs a two-stage classification process. First, the transcript is sent via Groq’s API to invoke [OpenAI’s GPT OSS](#), a 120 billion parameter MoE model built upon Generative Pre-trained Transformers (GPT), using GPT-2 and GPT-3 architectures [Vaswani et al., 2017, Radford and Narasimhan, 2018], optimized for complex reasoning tasks and efficient inference [Agarwal et al., 2025]. This model was chosen over rigid keyword matching because grocery queries show high linguistic variability: users may phrase the same intent in several ways (e.g., “Where are apples?”, “Guide me to apples”, “How do I reach apples?”), and a LLM’s few-shot reasoning capability generalizes better in these circumstances [Sakai et al., 2025].

The intent classifier instructs the model to classify the user’s intent using the input query into one of the five categories (four SLIV tasks and 1 fallback), described in Table 3.2.

The full prompt is reproduced below:

You are an intent classification assistant for a grocery shopping aid. Your goal is to analyze the user’s input and categorize it into one of five specific schema. Analyze the input step-by-step in your “reasoning” field before selecting the final “category” and “content”.

Classification Rules:

Table 3.2: Intent Categories, their corresponding SLIV tasks and triggers.

Intent	Task	User Query
Scene	Scan	Describe the immediate environment
Object	Locate	Query about a specific object
Guidance	Interact	Guidance to reach a specific object
Validation	Validate	Price, expiry date or other product details
Chat	Fallback	Intent could not be classified into the above categories, user query and image is sent to a simple VLM for response

1. **SCENE** — Trigger: User asks for a general description of the environment or their surroundings. Content: Leave empty string “”. Example: “What is around me?”, “Describe the view.”
2. **OBJECT** — Trigger: User asks if a specific item exists, is looking for something specific, or points something out. Content: Extract the specific object name (noun phrase). Example: “Do you have apples?”, “I am looking for milk.”, “What is that red thing?” → “red thing”
3. **GUIDANCE** — Trigger: User asks for navigation help or how to reach a specific item. Content: Extract the destination object. Example: “How do I get to the dairy section?”, “Where are the cookies?”
4. **VALIDATION** — Trigger: User asks for confirmation, validation or reassurance about a particular object. Content: Extract the object details. Example: “The apples are priced at \$3.99 CAD per pound.”
5. **CHAT** — Trigger: Any input irrelevant to grocery shopping, navigation, or visual queries. Content: Leave empty string “”. Example: “What is the capital of France?”, “Hello there.”

Output Requirement: You must populate the “reasoning” field first with your analysis, then the “intent” (one of: scene, object, guidance, validation, chat), and finally the “content”.

This structured-output design is motivated by prior work on schema-constrained generation. [Lu et al. \[2025\]](#) show that constraining a model’s output to a schema narrows its

generation space, improving accuracy on classification and extraction tasks by reducing both syntax errors and logical mistakes. Shorten et al. [2024] similarly report an 82.55% success rate for models producing correctly structured output. Thus, the prompt instructs the model to output a structured JavaScript Object Notation (JSON) object with three fields:

- *reasoning*: Step-by-step analysis of the user query (CoT reasoning to improve accuracy [Wei et al., 2022]).
- *intent*: One of {scene, object, guidance, validation, chat}.
- *content*: Extracted object name (for object description/guidance/validation) or empty string (for scene description/fallback chat).

The *chat* intent deals with edge cases where the query does not map cleanly to SLIV tasks (e.g., social pleasantries, meta-questions about the system). By explicitly modeling this category, ShelfScout avoids forcing ambiguous queries into inappropriate processing pipelines.

3.5.2 Visual Reasoning Workflows

Based on the classified intent, the n8n orchestrator routes the query to one of four visual reasoning pipelines or the chat fallback. Each pipeline is optimized for a specific SLIV task while employing the same VLM: Meta’s Llama-4 Maverick,¹ a 17 billion parameter VLM optimized for visual question-answering [Touvron et al., 2023], via a Groq API call.

This model was selected because unlike “late-fusion” models that process images through a separate encoder (e.g., OpenAI’s Contrastive Language-Image Pre-training (CLIP)) before passing them to the LLM, Llama-4 Maverick utilizes early-fusion multi-modality [Abdullah et al., 2025]. It allows the model to treat image and text tokens within a unified backbone from the first layer, preventing “semantic bottlenecks” that occur when an image is compressed into a single vector by a CLIP-encoder before the LLM sees it, improving its ability to ground specific text labels (e.g., “yogurt”) to precise spatial patches in the image [AI, 2025b]. Also, by leveraging Groq’s Language Processing Unit (LPU) inference engine, Maverick achieves throughputs exceeding 500 tokens per second [AI, 2025a]. This

¹This model is an instruction-tuned variant with extended context length.

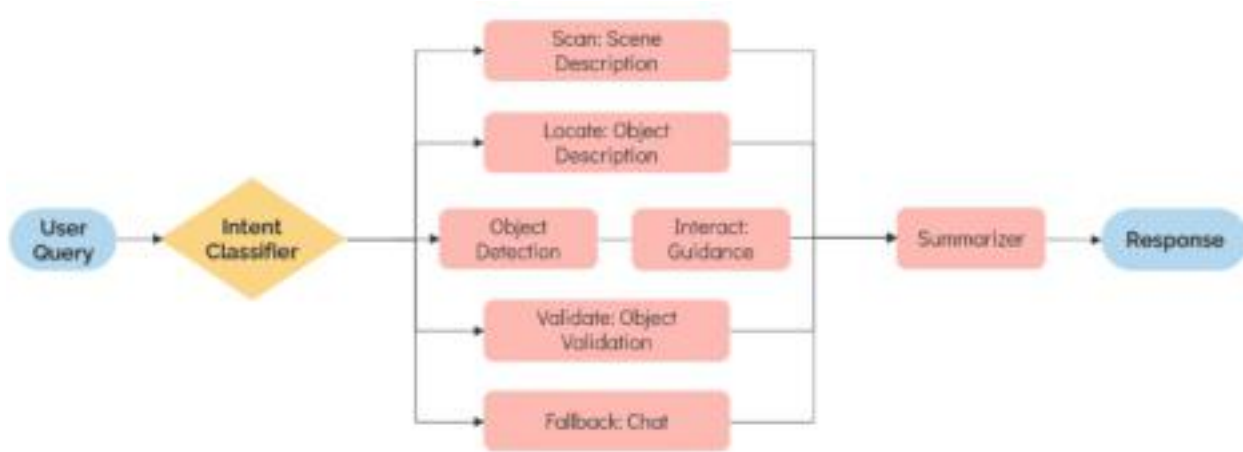


Figure 3.1: Visual Reasoning Workflows implementing each of the four SLIV tasks on Shelf-Scout’s backend.

near-instantaneous response time is crucial for assistive technologies where users require immediate feedback to adjust their orientation in real-time. Fig. 3.1 illustrates the routing logic.

1. Scene Description: For *intent=scene*, the model was asked to generate a high-level description of the user’s immediate surroundings.

You are helping a blind grocery shopper understand their environment. When asked to describe the scene, do not mention irrelevant things like lighting, floor or ceiling. Instead, describe the grocery layout in terms of aisles, shelves and products. Mention what is on the left, right or in front of the user (e.g., ‘The aisle to your right has cereal boxes’). Estimate approximate distance in steps or meters (‘the aisle is about 3 steps away or 3 meters away’). Mention whether there are obstructions or people nearby. If it’s an open shelf/section, describe its category (e.g., bakery, produce, frozen goods). Keep the explanation simple, concise and actionable for someone who cannot see. Do not add extra visual descriptions that are not useful for navigation or shopping.

2. Object Description: For *intent=object*, the model was simply asked to describe the object:

Your task is to look for the following object in the image: [object name] and then briefly describe it using descriptive language.

3. Guidance: For *intent=guidance*, the model was provided an annotated image with a bounding box around the target object, created by running [OmDet](#) locally for open-vocabulary object detection [Zhao et al., 2024a] via a “vision-pipeline” docker container, and was asked to generate step-wise instructions to guide the user to reach the object. The decision to feed the model with an annotated image around the object of interest was backed by previous work by Yang et al. [2023], who demonstrated that LLMs, when given annotated images, out-performed state-of-the-art fine-tuned models for grounding and spatial tasks. In order to ensure grounded guidance instruction generation, the model was specifically asked to “look” for the annotated object.

Your task is to provide the user with step-by-step instructions on how to reach the [object name] in this image. The image was annotated with bounding boxes and an arrow that indicates the object of interest. Provide instructions on how to reach the object. Use way-finding instruction words like ‘left’, ‘right’, ‘forward’, ‘backward’, ‘ahead’, ‘number of steps’, ‘at eye level’, ‘at ground level’, etc., to guide the user towards the object. Assume that the view you have from the image is the view through the user’s eyes.

It is important to note that this preliminary version of the system did not support guidance towards ‘out-of-frame’ objects, a limitation that was addressed by the incorporation of the RTAB-Map and ARKit guidance pipelines by colleagues, for which, the backend infrastructure will be described in detail in the final system design in Chapter 5.

4. Validation: For *intent=validation*, the prompt was as follows:

The user is looking for [object name]. Your task is to make sure the most prominent object² in this image is what the user is looking for. Then describe the object’s shape, colors and read any information present in the image.

²Note that if the user was looking for an object and it was not the “most prominent object”, the model was still able to identify and provide the location of the target object if it was visible in the background.

5. Chat: For *intent=chat*, the model was provided with the user’s query and the image and was not given any other instructions. It could respond to the user query as seen fit.

3.5.3 Summarizer

All outputs from the visual reasoning pipelines (scene description, object description, guidance, validation, chat) are passed through a final *summarization* step before being transmitted to the front-end. This ensures responses are concise and free of technical jargon that could overwhelm BVI users in grocery store environments [Gamage et al., 2023]. The summarizer is implemented as an additional Groq API call to OpenAI’s GPT OSS with a prompt that receives the raw JSON outputs and the original user query:

You are a helpful assistant for a BVI user. Your goal is to synthesize the provided technical visual data into a helpful, natural spoken response.

INPUT DATA: *You will receive raw JSON data describing the scene, objects or spatial relationships.*

TASK:

- Interpret the JSON data.
- Formulate a response.
- Adjust your tone and length.

RULES:

- Use natural language. No JSON or technical jargon in the output.
- If the visual data contains hazard warnings, prioritize them.
- If the visual data is low confidence, convey that uncertainty politely.
- Respond briefly and clearly.

The structured output from the summarizer is then sent to the front-end via Hypertext Transfer Protocol (HTTP) response, which synthesizes the text into speech, which is played immediately, completing the interaction loop.

Chapter 4

Preliminary User Study

Building directly upon the preliminary architecture of ShelfScout outlined in Chapter 3, this chapter presents the results from a response-focused user study, which was conducted from August 29, 2025 to September 6, 2025, designed to expose early failure modes before physical deployment. Instead of evaluating end-to-end physical usability, this phase isolated and benchmarked the quality of the system-generated responses against established commercial AI platforms. The preliminary study was designed with three primary goals:

1. To identify which technical aspects of the system needed additional “safeguards” or architectural restructuring (BVI-friendly response phrasing, hallucination-mitigation strategies, navigation support).
2. To understand what information BVI users prioritize in assistive scenarios (conciseness versus detail, spatial risk cues, product attributes such as price, expiry date, brand).
3. To contextualize the prototype’s performance by comparing it to three commercial AI systems, namely, Be My AI, Meta Ray-Ban smart-glasses and Gemini Live.

The systems were evaluated through an audio-only remote session with BVI participants, who provided their feedback regarding the clarity, trustworthiness, usefulness and perceived correctness of pre-recorded system-generated responses, factors identified as critical for assistive technology adoption by [Gamage et al. \[2023\]](#) and [Kuriakose et al. \[2020\]](#), highlighting important aspects of information adequacy and response quality that shaped the final design of ShelfScout described in Chapter 5.

4.1 Baselines for Comparison

ShelfScout was compared to three other vision-based AI systems representing different interaction paradigms and underlying model families:

- **Be My AI:** As a widely used GPT-4V application, Be My AI is currently the “de facto” standard, serving as a strong baseline due to its real-world deployment and high familiarity within the BVI community [Zhao et al., 2024b], making it an appropriate benchmark for assessing whether ShelfScout offers improvements in hallucination control, spatial grounding and shopping-specific guidance rather than only matching generic description quality. It represents the “chat-based” paradigm, where users send a photo and receive a detailed paragraph in return.
- **Meta Ray-Ban Smart-Glasses:** A hands-free wearable system integrating cameras and microphones into glasses, powered by Meta’s Llama-based MLLMs. This baseline represents the “wearable computing” paradigm where users interact through voice commands and tap gestures without needing to hold a smartphone. Comparing ShelfScout to this mainstream wearable allows an examination of the trade-offs between a vertically integrated “black box” vendor system and the proposed modular architecture.
- **Gemini Live:** A real-time conversational system that continuously captures audio and visual input and maintains an ongoing dialogue rather than discrete turn-based prompts. This baseline represents the “continuous input” paradigm, where the model can monitor context over time and generate follow-up questions¹ or clarifications without explicit user re-prompting.

These systems were chosen to span the spectrum of current VL assistance approaches: chat-based interaction (Be My AI), wearable computing (Meta Ray-Ban), real-time continuous input (Gemini Live) and task-decomposed approach (ShelfScout), each built on different foundation models and deployment strategies. This diversity enables assessment of whether ShelfScout’s architectural decisions— mixture of locally hosted and API-served

¹For example, a user might ask, “I’m looking for some breakfast items. Can you guide me to them?” The system could respond, “There are several cereal boxes to your right. Which breakfast items, in particular, would you like to find?”, generating a follow-up question to clarify the user’s request.

LLMs/VLMs, explicit task routing and SLIV-based decomposition—yield perceived benefits over existing commercial applications.

4.2 Experimental Setup

To generate the pre-recorded audio responses used in the user evaluation, all four systems were tested in a real-world grocery store during regular operating hours, with permission from the store management. Researchers used standardized prompts derived from realistic scenarios and linguistic patterns of BVI queries observed in the VizWiz dataset [Gurari et al., 2018] to query each of the four systems and record the generated responses.

4.2.1 Aisles

The evaluation was conducted across four grocery store aisles, each presenting distinct visual challenges:

- **Produce:** Contains fresh fruits and vegetables with varied shapes and colors. These test the system’s ability to identify organic products without relying on packaging or text.
- **Dry Goods:** Features boxed and canned items with textual information and brand identifiers. This environment evaluates text recognition capabilities and product differentiation among similar items.
- **Bakery:** Contains baked goods such as bread, pastries, and rolls, often in translucent or opaque paper bags with non-standardized printed labels. Unlike produce, bakery items typically have some form of labeling, though the print quality, placement, and legibility can vary significantly. Hence, this tests the system’s ability to recognize semi-packaged goods and interpret text-based cues that are less uniform than those found on standardized products.
- **Cold Section:** Contains cold beverages and dairy products, sometimes with condensation on the packaging. This environment tests performance under challenging lighting conditions and reflective glass refrigerator doors.

4.2.2 Tasks and Prompts

For each aisle-system combination, the task of finding the correct object was broken down according to the SLIV framework and corresponding prompt:

1. **Scene Description:** “Describe the scene in front of me.”
2. **Object Description:** “Which one is the [target object]?”
3. **Guidance:** “Can you please guide me to the [target object]?”
4. **Validation:** “Can you please describe this label?”

The target objects included cauliflower, Kellogg’s Rice Krispies, pepperoni pizza and vanilla or strawberry cupcakes.²

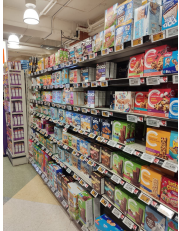
All four prompts remained identical across systems for a given aisle, ensuring that differences in responses reflected system capabilities rather than query variation. For example, the validation task consistently used the phrasing “Can you please describe this label?” and did not include alternatives such as “Can you read the label?” or “Can you give me product-relevant information for this?”, meaning that this study did not assess how robust or reliable each system’s responses were to natural linguistic variability in user queries.

Each environment was photographed using two types of images: overview shots (captured at chest-height) for scene description, object description and guidance tasks; and close-up views (at arm’s length) for object validation. This approach simulated how a BVI user might naturally capture their surroundings, first orienting within the broader environment before focusing on specific items. While the Meta Ray-Ban glasses captured slightly different perspectives due to their head-mounted form factor, we maintained approximately similar user positioning and camera orientation. The dataset of images with their corresponding prompts is illustrated in Table 4.1.

Each trial followed the system’s native interaction paradigm: Be My AI required image capture followed by chat interface interaction; Meta Ray-Ban glasses responded to tap gestures or voice triggers; Gemini Live used voice-only input after initial setup; and ShelfScout

²Vanilla cupcakes were selected as the target object. However, on one of the data collection days, the grocery store did not stock vanilla cupcakes, so strawberry cupcakes were used as a substitute.

Table 4.1: Images used for evaluation across grocery store sections with their corresponding tasks and prompts.

Task & Prompt	Overview Images			
Scene Description: “Describe the scene in front of me.” Object Description: “Which one is the [target object]?” Guidance: “Please guide me to the [target object].”	 Produce goods.	 Boxed products such as cereal and granola bars.	 Cupcakes, donuts and other baked goods.	 Refrigerators containing frozen pizza boxes.
	Produce	Dry Goods	Bakery	Cold Section
Validation: “Can you please describe this label?”	 Cauliflowers in the produce section.	 Rice krispies in the dry goods section.	 Cupcakes in the bakery section.	 Pizza boxes in the cold section.
	Close-up Images			

required two screen taps for starting and stopping audio recording, while automatically taking pictures. This protocol yielded 64 combinations (4 sections $\times$ 4 tasks $\times$ 4 systems), with all responses saved as text files for speech synthesis.

4.3 Remote Study Design

To efficiently gather feedback from BVI individuals from different locations, participants joined [Zoom](#) sessions where they listened to pre-recorded system responses and provided structured feedback via questionnaires and semi-structured interviews. This methodology prioritized response quality assessment over an end-to-end usability test. Testing four systems across four aisles in a single 2-hour session would be infeasible in physical stores due to travel logistics and participant fatigue. Remote sessions required no transportation to grocery stores, reducing social anxiety and logistical barriers documented in prior BVI shopping research [[Khattab et al., 2015](#)].

4.3.1 Study Protocol

A pilot session with one blind participant validated the procedure, prompt phrasing and questionnaire structure before proceeding with the full study. Each participant attended a single 90-120 minute Zoom session structured into four phases: introduction and pre-test questionnaire, system response evaluation, post-aisle questionnaires and a semi-structured comparative interview. Upon joining, participants were read a standardized introductory script that explained the study’s purpose, voluntary nature and compensation. After obtaining informed consent for both participation and audio recording, the participant was presented the core evaluation framework:

This study is about evaluating different AI systems designed to support blind and low vision shoppers in grocery stores. We want to understand how helpful and usable these systems are when completing common shopping tasks. In this session, we’ll ask you to imagine that you are in a grocery store. As you move through different aisles, you’ll hear pre-recorded audio responses generated by four different AI systems to pre-defined queries, running on a smartphone. Each system has been given the same set of four tasks to complete across different grocery sections such as produce, bakery, frozen goods or dry goods.

Participants were explicitly instructed that they would hear responses from all four systems but that each would be assigned to a different aisle from a grocery shopping experience,

ensuring exposure to varied product contexts (e.g., produce vs. packaged goods) while avoiding repetition bias. Participants were *not* informed which specific systems they were hearing (ShelfScout, Be My AI, Meta Ray-Ban, Gemini Live), nor were they told the order in which the systems appeared, to minimize expectation bias. To eliminate voice-based bias, all system responses were synthesized using Kokoro TTS [hexgrad, 2025] before being presented to participants. This standardization ensured that participants evaluated response *content* (accuracy, verbosity, clarity) rather than voice quality. Each audio clip lasted approximately 10-15 seconds, though some systems produced longer responses for the same prompt, which itself became a point of feedback.

The evaluation followed a counterbalanced Latin square design where each participant encountered all four systems, each system assigned to a different aisle, to control for order effects and aisle-specific difficulty (e.g., ensuring that no single system was consistently evaluated in the visually simpler produce section versus the cluttered dry goods aisle). For each aisle-system evaluation, a standardized protocol was followed:

- **Context Setting:** The aisle type was verbally described to establish the imagined environment (e.g., “You are now in the frozen goods/refrigerated section.”).
- **Task Presentation:** Before playing each audio response, the corresponding prompt was read aloud exactly as it would have been spoken to the system for real-world testing (e.g., The prompt you gave the system was: “Describe the scene in front of me”, followed by playback of the system’s audio response).
- **Immediate Feedback:** After each audio clip, participants were invited to share initial reactions, noting aspects they liked or disliked about the response. Participants were explicitly told they could choose not to comment on every audio if they had no strong reaction, and could request replays if needed. They were also encouraged to adopt a “Think Aloud” protocol, verbally articulating their immediate impressions and cognitive processes. They were invited to share any observations, even those that were tentative or poorly defined, regarding specific elements that intrigued them or caused dissatisfaction with a response.
- **Post-Aisle Questionnaire:** After hearing all four task-responses from one system within a single aisle, participants completed a brief structured questionnaire, rating

the system on several dimensions:

- **Confidence:** Extent to which participants reported feeling capable of completing tasks independently with the aid of the system (“Are you confident in your ability to perform the task using the system-generated responses?”).
- **Trustworthiness:** Perceived trust in the system’s ability to provide accurate and reliable information (“Would you trust this system’s information to make shopping decisions?”).
- **Actionability:** Defined as the extent to which participants felt that the systems’ responses provided sufficient guidance to complete the task (“Did the responses provide sufficient information to complete the intended task?”).

This sequence repeated for all four aisles, exposing each participant to all four systems. Sessions included participant-requested breaks to mitigate fatigue. Upon completing all system-aisle evaluations, participants engaged in a semi-structured interview lasting approximately 5-10 minutes, wherein they were asked to rank the four systems according to their overall preference and identify which they found most trustworthy and natural-sounding. Participants were probed for explanations regarding their feedback. Participants were also asked to suggest improvements or features that they would prioritize for an ideal assistive shopping system. All interviews were audio-recorded (with consent) and later transcribed for qualitative analysis.

4.3.2 Ethics and Recruitment

Participants were recruited under ethics approval from McGill University’s Research Ethics Board (REB) (File Number: 25-06-021). Recruitment targeted blind and low-vision adults with experience using mainstream assistive technologies, reflecting the intended end-user population for ShelfScout.

During the study, the researcher read the consent form aloud and obtained verbal consent to participate and to record audio for research purposes. Participants were reminded that participation was voluntary, that they could pause or withdraw at any time without penalty and that their data would be anonymized in all reporting. Compensation was set at CAD 20 per hour, pro-rated to the nearest 15-minute interval.

A total of 8 BVI adults completed the remote study. Participants ranged in age from 28 to 60 years (mean age = 51 years). The sample was gender-balanced, comprising four men and four women. Most participants reported the onset of vision impairment in early childhood (≤ 12 years), with one participant reporting adult-onset progressive vision loss. Vision status covered both total blindness and severe/legally blind low vision.

Pre-test questionnaires collected detailed information about each participant’s everyday life and technology use. Participants reported whether they relied on human assistance for navigation and daily tasks, accompanied by explanatory comments (e.g., relying on relatives after moving to a new city, complementing guide-dog use with occasional human assistance). This information contextualized how strongly participants currently depend on human helpers versus technological aids. Given the study’s focus on AI-powered visual assistance, the questionnaire probed current and previous use of Vision Assistance Technologies (VAT). Participants indicated which tools they used from a list including Be My Eyes, Be My AI, Seeing AI, Aira, smartphone screen readers (e.g., VoiceOver on iOS), navigation apps such as VoiceVista, Soundscape, BlindSquare and various smart-glasses platforms (e.g., Meta/Ally glasses). Several participants described specific pain points such as needing to take multiple photos to extract the right information, long or overly verbose descriptions that obscured key details such as price, or concerns about hallucinations and inconsistent answers. Others emphasized the strengths of existing systems (e.g., effective signage reading when in the correct aisle, strong utility of smart-glasses for rapid queries).

4.4 Study Results

The evaluation of the system-generated responses was categorized into a “researcher-led” analysis and a “participant feedback” qualitative analysis. The researcher-led evaluation consisted of examining whether system responses aligned with ground truth characteristics of the scene. Any deviations were categorized as hallucinated details, encompassing both fabricated information (e.g., invented objects mentioned that were not present in the real-world scene) and inaccurate information (e.g., incorrect counts, mislabeled brands, or wrong prices). Responses were marked as either free of hallucinations (✓) or containing fabricated details (×). In addition, the *hallucination rate* of each system was calculated by dividing the total number of hallucinations by the total number of words across all responses for each

system. This served as an estimate of how much of the generated content from each system is fabricated. The participant feedback analysis highlighted the clarity, trustworthiness and actionability of the responses for BVI grocery shopping.

4.4.1 Researcher-Led Evaluation

Table 4.2 captures whether system responses contained hallucinated visual elements such as fabricated objects, inaccurate locations or misread labels.

Table 4.2: Summary of hallucination avoidance for each system across grocery store sections and tasks. A check mark (✓) indicates that the system avoided hallucinating visual elements in its response. A cross (×) indicates that the response included some elements that were fabricated or not entirely accurate.

System	Task	Produce	Dry Goods	Bakery	Cold Section
Be My AI	Scene Description	×	✓	×	✓
	Object Description	✓	✓	×	✓
	Guidance	✓	✓	×	✓
	Validation	×	✓	×	✓
Meta Ray-Ban	Scene Description	×	✓	✓	✓
	Object Description	×	✓	×	×
	Guidance	×	✓	✓	×
	Validation	✓	×	✓	✓
Gemini Live	Scene Description	✓	✓	✓	✓
	Object Description	✓	✓	×	✓
	Guidance	✓	✓	×	✓
	Validation	✓	✓	✓	✓
ShelfScout	Scene Description	✓	✓	✓	✓
	Object Description	✓	×	✓	✓
	Guidance	✓	✓	✓	✓
	Validation	×	✓	×	✓

Be My AI was the most verbose system (averaging 56 words per response) and had a low hallucination rate, recording 0.007 hallucinations per output word. Meta Ray-Ban glasses produced brief responses (averaging just 25 words per response). It also had the highest hallucination rate per output word (0.03). Gemini Live and ShelfScout had shorter

responses (averaging 29 and 35 words per response, respectively) and fewer hallucinations overall (average of 0.004 and 0.007 hallucinations per output word, respectively). A task-specific analysis of the four systems yielded the following observations:

Scene Description

All systems demonstrated strong capabilities in describing the environment. Be My AI tended to be more verbose and, in one instance, introduced minor hallucinations (i.e., mentioning mushrooms that were not present in the produce aisle). Meta Ray-Ban glasses suffered from the same problem, but more frequently. Still, it was able to capture spatial layouts well. Gemini Live produced mostly accurate and grounded descriptions. ShelfScout provided slightly more concise and focused descriptions (35 words per response, 0.007 hallucinations per output word), aiming to maintain a balance between accuracy and brevity.

Object Description

All four systems were generally accurate in identifying the correct object when prompted. Be My AI was more likely to rely on surface text (OCR style reading), as in reading out a product bar code in one instance. Meta Ray-Ban glasses needed to be prompted twice in the produce and cold sections, and hallucinated product variants and shelf levels. Gemini Live identified the presence of the required objects but gave very brief responses, and was thus found lacking in relevant details. For example, in the dry goods aisle, it stated: “Okay, I see a few different brands like Kellogg’s Rice Krispies and General Mills Dunkaroos. Are you looking for a specific size or flavor?”—providing high-level brand recognition but omitting crucial product-specific attributes such as package color, price, or shelf placement. In contrast, ShelfScout was able to provide detailed spatial and descriptive information absent from Gemini Live’s output for the exact same task (e.g., “The Kellogg’s Rice Krispies are the blue and orange boxes on the third shelf from the top. They are placed prominently there.”). In addition, it showed improved grounding in the visual context when dealing with items differentiated by subtle visual cues (e.g., distinguishing strawberry cupcakes from plain cupcakes, where only a small red frosting swirl and label text differentiated the target item).

Guidance

Be My AI sometimes deferred to human support (e.g., in the Produce section, it stated: “I can describe the location, but I can’t physically guide you. . . please consider pressing the “Call a volunteer” button for guidance.”) or gave descriptive locations without step-wise instructions. Meta Ray-Ban glasses frequently described object locations incorrectly. For example, in one instance, it mentioned that a product was on the top shelf, while it was actually located on the third shelf from the top; while in another, it refused to provide guidance instructions altogether. Gemini Live failed to provide stepwise guidance instructions in most cases (e.g., in the Cold Section, it responded to the guidance query by saying, “Sure, the pepperoni pizzas are in the orange boxes on the bottom right. You can see the word “Pepperoni” on the front of the box.”). ShelfScout offered more spatially grounded guidance compared to the other systems (e.g., “step forward, turn slightly left”).

Validation

Be My AI misread price tags and reported package counts incorrectly. Meta Ray-Ban glasses provided short but accurate responses to the prompts. However, the responses were lacking in details such as product descriptions, price tags, product weight and expiry date. It was also the only system that required repeated prompting in the bakery section to produce a successful response. ShelfScout avoided some of these errors but introduced others, such as confusing an image of a cupcake for a strawberry. Gemini Live was the only system that remained hallucination-free in all validation responses but lacked descriptive details about the target object.

4.4.2 Participant Feedback

To ensure that all study materials were accessible non-visually, questionnaires were administered orally during the Zoom sessions. The researcher read each question aloud, including all response options for Likert-scale and multiple-choice items, and participants provided their answers verbally. The researcher then recorded these responses directly into the digital questionnaire (implemented using Microsoft Forms) on the participant’s behalf, effectively acting as a scribe while preserving the participant’s control over all choices. Participants were explicitly encouraged to request repetition of any question or set of options as many

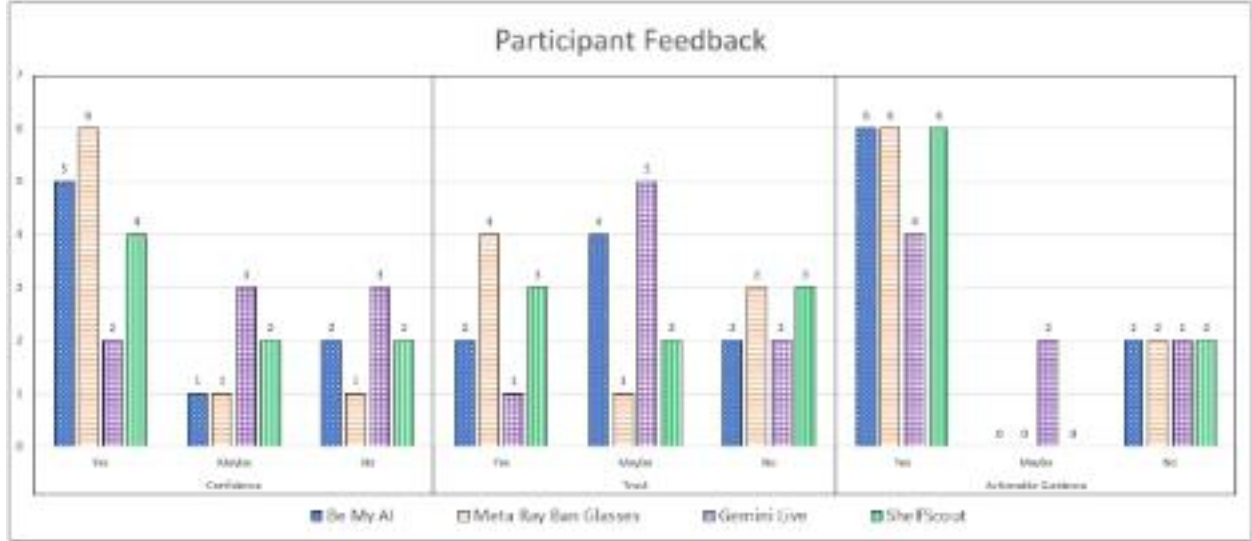


Figure 4.1: BVI participant feedback rating each of the four systems’ overall responses with respect to perceived confidence, trust and actionable guidance.

times as needed, and questions were paced to match each participant’s preferred speed of interaction. Open-ended questions were handled similarly: prompts were read aloud, participants responded in their own words and the researcher transcribed their responses verbatim, ensuring that participants could fully engage with the study without needing to manipulate visual forms or on-screen controls.

Questionnaire Feedback

Findings from the post-aisle questionnaires are summarized below and illustrated in Fig. 4.1.

- Meta Ray-Ban glasses received the strongest preference, with most participants expressing confidence in using this system to complete tasks independently.
- Meta Ray-Ban glasses were also considered most trustworthy across all test scenarios.
- All systems were found to provide actionable guidance in most aisles, with Gemini Live scoring slightly lower than the others.

Qualitative Feedback

Participants’ open-ended feedback from both the post-aisle questionnaire as well as the semi-structured interview was summarized to identify common themes.

Scene Description: Although most participants were satisfied with the scene description responses, some suggested that including information about which aisles lie to their left and right, along with an overview of the products located in those aisles, would provide a clearer sense of orientation within the store.

Redundancy in Object Description and Guidance Responses: Many participants expressed a preference for object description and guidance tasks to be combined into a single set of instructions that inform whether the target object is near and then provide stepwise navigation cues. This matched their mental model of scanning their environment to locate the desired object in real-world scenarios. They mentioned how current object description responses simply confirmed the presence of the target object on the shelf whereas they would have preferred to know “where on the shelf” the object is. For instance, a participant evaluating Gemini Live in the Bakery aisle stated, “Guidance was redundant as the system’s response sounded too similar to object description. Not helpful.” Hearing two separate responses in our study setup made the tasks feel unnecessary and repetitive, even though the underlying functionality would be essential in a real shopping task that was not “directed” by our study protocol.

Lack of Actionable Guidance Instructions: The usefulness of orientation cues was emphasized repeatedly. Participants desired clear egocentric references about product position, such as “to your right,” or “directly in front of you,” to better support spatial awareness. Others highlighted the value of reference points from the environment, such as shelf levels (“on the second shelf from the bottom”) or spatial anchors such as the end of an aisle. Finally, participants noted the importance of object-to-object references, where the target item is described in relation to nearby items (e.g., “next to the carrots,” or “to the left of the oatmeal box”). Several participants also mentioned that spatially relative instructions such as “bottom right” were hard to interpret in practice: one participant testing the Meta Ray-Ban glasses in the Bakery aisle noted: “second row from right: what does that even

mean?! Is it the shelf? Didn't make sense!" Instead, they preferred guidance that combined multiple reference frames, for example, instructions given with respect to "landmarks" such as surrounding objects to the target object and shelf levels. One participant stated that good guidance would follow a structure similar to "you are facing this product and to its right is the one you asked for..." This structured approach was perceived as more intuitive and actionable for supporting independent navigation, yet it was absent from the responses of all tested systems, including ShelfScout.

Color Descriptions: A recurring criticism concerned the reliance on color-based descriptions. While color-related details aided those with low vision in terms of guidance and validation, fully blind participants found them confusing or even "useless". One participant reflected on color-focused object detection from Gemini Live in the Produce aisle: "I can't see! Color information is useless to me! I don't need to know what color the shelves are." In addition, participants mentioned that highlighting the shape or size of the product would be more helpful than simply focusing on the product visuals such as graphics on the packaging.

Relevant Information during Validation: On this task, feedback involved a preference for specific details such as price, weight and product shape over packaging design or information such as bar codes. One participant expressed frustration with Be My AI in the Dry Goods aisle: "Please don't read me the bar code! I just want to know how big it is and how much it costs!" Participants also highlighted that most systems failed to mention details such as product variants, ingredient details including presence of allergens, quality of selected product (e.g., whether the produce is moldy or the packaging is dented), as well as availability of cheaper or discounted products.

Response Language and Presentation Style: Several users were confused or frustrated by unsolicited translations of bilingual product labels while being fluent in both languages—e.g., "Why is it translating to me? Personally, I'm very fluent in French. It's a bit of waste of time if you haven't really requested translation from French to English." Also, some systems emphasized products with colorful packaging during scene descriptions, giving participants the impression that their attention was being forcefully drawn to specific products or brands. Concerns were also raised about perceived bias in Gemini Live's responses in the

Cold Section, including the impression of brand favoritism (“Why did it pick out specific brands. . . sounds like the system wants to control the information it gives you”). Finally, articulation choices impacted usability, such as stating phrases such as “I see the label” during validation, which participants deemed redundant or unnecessary. On the other hand, some participants valued human-like conversational patterns, such as Gemini Live’s final remark, “Are you looking for anything else?”, which positively impacted engagement.

4.4.3 Limitations of the Study

The remote audio-only format prevented evaluation of system latency, physical reaching accuracy or the real-time cognitive and motor demands of navigating a crowded store while using the AI system on a smartphone. Also, evaluating the prototype using single, fixed prompt templates across all aisles meant the study could not assess system robustness against variations in query paraphrasing. Lastly, the sample size of eight participants limited generalizability but was prioritized to capture in-depth qualitative insights to guide system refinement.

Chapter 5

System Evolution

Building on the insights from the initial Zoom-based user study and subsequent internal testing, the ShelfScout architecture underwent several iterations. This chapter presents the final backend architecture developed upon the foundational system design outlined in Chapter 3.

5.1 Architectural Changes

The final ShelfScout architecture differs from the preliminary design in four key ways:

- **Introduction of a Presence Check Module:** Early testing revealed that without object presence verification, the system would generate guidance toward non-existent items (Table 4.2). A verification step was therefore added before routing to downstream visual reasoning pipelines, ensuring the system does not guide users toward products not actually present in their immediate environment.
- **Adoption of Open-Vocabulary Detection via Qwen2.5-VL:** Internal testing revealed that OmDet struggled with ambiguous object categories (e.g., “breakfast items”) and visually similar products (e.g., identifying specific cereal brands). The system replaces this with a locally hosted instance of [Alibaba’s Qwen2.5-VL](#)¹ [Bai et al., 2023], enabling more robust open-vocabulary detection across diverse grocery items.
- **Decoupling of Guidance Pipelines:** Preliminary testing showed that a single guidance pipeline could not smoothly transition between aisle-level navigation and object-

¹This model has four billion parameters and is instruction-tuned.

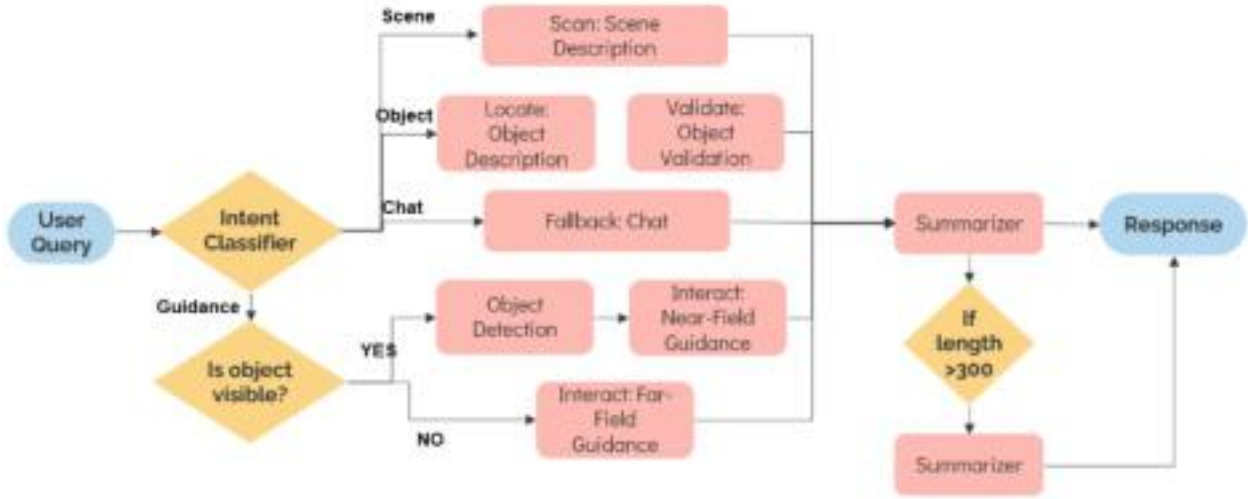


Figure 5.1: Updated Visual Reasoning Workflows implementing each of the four SLIV tasks on ShelfScout’s backend.

level reaching. Separating guidance into far-field (store navigation) and near-field (hand-level reaching) pipelines enables continuous end-to-end assistance, addressing design goal 2 (sequential guidance), stated in Section 3.1.

- **Refined Multi-Stage Reasoning Pipelines:** Users indicated redundancy in having to prompt the system for object description and guidance tasks separately (Section 4.4.2). The workflow was refined to automatically chain stages when appropriate. For example, a guidance query with a visible object would trigger high-level object location description first, followed by the far-field navigation and near-field reaching instructions in sequence without user re-prompting, maintaining continuity across the four tasks.

The updated architecture highlighting the modifications made to the visual pipelines is depicted in Fig. 5.1.

5.1.1 Intent Classification

The intent classification component retains the prompt structure introduced in Section 3.5.1. However, it was simplified to better align with the revised pipeline architecture. The origi-

nal five-way classification scheme was reduced to four categories by removing the dedicated *validation* intent. Instead, all queries related to object identification, description or verification are now uniformly classified under the *object* category.² This consolidation of the object-related reasoning tasks into a single pipeline reduces ambiguity in routing decisions and simplifies downstream processing.

5.1.2 Object Presence Check

For queries classified as *guidance*, ShelfScout performs an additional verification step before invoking visual reasoning pipelines. This “presence check” determines whether the target object is visible in the current camera frame, enabling the system to route the query to either:

- **Near-field guidance:** If intent is classified as *guidance* (indicating guidance phase) and the object is present in frame, the system triggers the ARKit-based hand-level guidance pipeline that provides the “last-meter” guidance instructions for the user to reach for and grasp the required object. This guidance pipeline was developed by a colleague, as mentioned in Section 1.3, and will hence be treated as a “black-box” for this thesis.
- **Far-field navigation:** If intent is classified as *guidance* and the object is absent from the frame, the system triggers the RTAB-Map pipeline to locate the required object elsewhere in the store and provide aisle-level directions.

The presence check is implemented via a Groq API call to Llama-4 Maverick. The model receives the extracted object name and the camera frame, and is prompted to return a structured JSON response with two fields:

- *evidence*: A single factual sentence describing what is observed (or explicitly stating the object is not visible).
- *visible*: Boolean (*true/false*) to indicate object presence/absence.

²Since object description and validation are no longer classified separately, any query classified as *object* triggers both pipelines sequentially. The system returns a combined response containing both a high-level object description and fine-grained details such as price and expiry date, synthesized by the summarizer, eliminating the need for the user to re-prompt for each type of information separately.

The prompt emphasizes conservative behavior to prevent hallucinated presence claims:

You are a precise vision assistant for blind or visually impaired shoppers. You MUST respond with valid JSON matching this EXACT schema: {"evidence": "one short factual sentence describing what you observe or explicitly stating the object is not visible", "visible": true or false}. Examples: {"evidence": "A yogurt container is visible on the middle shelf, center-left.", "visible": true} or {"evidence": "No yogurt is visible in the image.", "visible": false}. If the object is found, describe its location. If NOT found, state explicitly it is not visible and set visible to false. Output ONLY this JSON - no other text.

By enforcing a binary decision with explicit evidence, this check reduces the risk of guiding users toward non-existent objects, a failure mode documented in prior VLM deployments [Liu et al., 2024].

5.1.3 Visual Reasoning Workflows

While the SLIV-based decomposition remains unchanged, the internal design of the visual reasoning workflows was refined to address limitations observed during evaluation. These updates focus on how queries are routed and processed within existing pipelines.

Note that due to Llama-4 Maverick no longer being supported on the free tier of Groq, all visual reasoning queries are now routed to *Meta’s Llama-4 Scout* instead.

Scene Description

For *intent=scene*, the prompt is modified to instruct the model to focus on navigable space, hazards and layout:

You are a vision assistant analyzing an environment. Focus on extracting navigable space, major obstacles and layout data.

Output Format: *Return a JSON object with exactly these keys:*

- *layout_data*: List visible pathways, aisles, or open spaces with approximate dimensions if possible.

- *hazard_data*: List immediate hazards (boxes on floor, people, poles) with location.
- *environment_summary*: A dense, factual summary of the room’s content.

This pipeline prioritizes orientation and safety over granular product details, aligning with the phase’s goal of environmental awareness [Gamage et al., 2023].

Far-field Guidance

When *intent=guidance* and *presence=false*, meaning the object is not visible in frame, Shelf-Scout invokes the far-field RTAB-Map-based guidance pipeline. The workflow is as follows:

1. **Localization in Database:** The backend sends the current camera frame and target object name to the RTAB-Map API endpoint */search-and-localize*. RTAB-Map checks whether the “target_object” exists in its own predefined database of objects and returns a *success* boolean (true: object exists in database, false: object does not exist). It also returns a “response” field with a single guidance instruction (e.g., “Walk 3 meters and then turn left.”)
2. **Guidance Generation:** If *success=true* and *response* is not null, the backend forwards the generated response to the front-end along with a “navigation=True” flag. This flag indicates that the system should continue requesting updated instructions as the user moves. The front-end repeatedly queries the backend, which directly calls RTAB-Map, bypassing earlier pre-processing steps to retrieve the next instruction, allowing the user to be guided step-by-step toward the target object. When the user reaches the vicinity of the object— within 1–2 meters and facing it— RTAB-Map returns a “reached=True” flag. The backend monitors this flag and, once received, informs the user that the object has been reached. At this point, the backend stops the navigation process, resets the “navigation” flag to *False* and triggers the near-field guidance pipeline. The transition happens automatically so that the user does not need to issue a new command, making the guidance experience continuous from navigating the store to reaching for an object.
3. **Fallback for Missing Objects:** If RTAB-Map returns *success=false* OR *response=null*, indicating that the object is either not in the database or the user could not be local-

ized, ShelfScout triggers a fallback pipeline. The system makes a VLM call to confirm the object’s absence:

You are a vision assistant looking for a missing item. The primary detector failed. Look for edge cases or partial visibility.

Output Format: *Return a JSON object with exactly these keys:*

- *scan_analysis*: Broad description of the image content.
- *candidate_objects*: List of any items that **might** be the target, with locations.
- *search_hints*: Objective suggestions for where to look next based on visual clues (e.g., “darker area to the right”).

This forces the system to admit uncertainty, prompting the user to reorient or switch to the *chat* fallback for exploratory queries (e.g., “Where else might I find apples?”).

Near-Field Guidance

When *intent=guidance* and *presence=true*, the pipeline assumes the user is close to the object, since the object is visible in frame, and needs fine-grained reaching instructions. This pipeline proceeds in three steps:

1. **Bounding box generation:** The image and object name are sent to a locally hosted Qwen2.5-VL container [Bai et al., 2023], which performs open-vocabulary object detection and returns bounding box coordinates (*x_min*, *y_min*, *x_max*, *y_max*) in pixel space. The model is prompted to output only valid JSON:

*You are a precise object detection assistant. Analyze images and return ONLY valid JSON with bounding boxes as absolute pixel coordinates [*x_min*, *y_min*, *x_max*, *y_max*]. Never include explanations, markdown, or extra text outside the JSON block. Always return complete JSON even if no objects found. Locate all visible instances of the target_object ‘object’ in this image. Return only a JSON block in this format: “objects”: [{"label": “target_object”, “bbox”: [*x_min*, *y_min*, *x_max*, *y_max*]}]. The bbox coordinates should be absolute pixel values, not normalized.*

The Qwen2.5-VL model replaces the previously used OmDet-based detection pipeline because the latter failed to detect ambiguous object names (e.g., “Find me breakfast items.”) and gave very low confidence scores for visually similar items during repeated internal evaluations.

In contrast, Qwen2.5-VL is trained for open-vocabulary detection and text-conditioned spatial grounding, allowing a direct mapping from free-form textual queries to precise bounding box coordinates [Bai et al., 2025]. This improves robustness in scenarios where object categories are under-specified (e.g., “breakfast items”), as the model can infer relevant visual instances without requiring predefined class labels. Also, the model shows strong fine-grained visual discrimination capabilities, allowing it to distinguish between visually similar products such as different flavors or variants within the same brand [Meng et al., 2024]. This is important in grocery shopping where visual differences are granular but meaningful.

Qwen2.5-VL outputs bounding box coordinates in structured JSON format. Because the model uses integer coordinates in a native $[0, 1000]$ reference frame rather than the camera’s true resolution, a lightweight linear scaling step rescales the bounding boxes to match the original image dimensions. This design reduces post-processing complexity and improves system latency, resulting in a more reliable and efficient pipeline for real-time guidance [Wang et al., 2024].

2. **Image annotation:** The bounding boxes are overlaid onto the original image using [OpenCV](#), producing an annotated frame with colored rectangles highlighting all instances of the detected object. This annotated image serves as a spatial grounding mechanism for the VLM in the next step.
3. **Spatial guidance generation:** The annotated image is sent to the VLM with a prompt that uses the visual annotations to produce high-level initial guidance:

You are a navigation assistant. The input image has bounding boxes drawn around detected objects. Use these annotations to extract accurate spatial data.

Output Format: *Return a JSON object with exactly these keys:*

- *spatial_relationship*: Objective description of position relative to the camera/user (e.g., “Bottom right, 2 meters away”).
- *reference_objects*: List of visible landmarks and objects nearby.

The resulting high-level guidance is returned to the user as an initial spatial cue. Following this, the backend issues a “reaching_ios” flag to the front-end, signaling the transition to the ARKit-based fine-grained guidance loop. Along with this flag, the system sends the object name, bounding box coordinates and the annotated image. The front-end uses this information to generate continuous step-wise navigation and hand-alignment instructions, guiding the user to grasp the target object.

Object Description and Validation

When *intent=object*, ShelfScout automatically triggers *both* object description and validation queries to minimize interaction overhead as well as providing the user with both high-level product features as well as fine-grained details in a single response, preventing them from re-prompting the system for each one separately. First, the system calls the VLM with a prompt designed to extract spatial context:

*You are a vision assistant locating a specific object. Describe its **context** and **spatial relationship** objectively.*

Output Format: *Return a JSON object with exactly these keys:*

- *location_data*: Detailed position info (shelf height, relative position).
- *surrounding_landmarks*: List of 1-2 prominent nearby objects.
- *visibility_factors*: Lighting, occlusion, or contrast notes.

Immediately after, a second call requests fine-grained validation details:

You are a meticulous product verification assistant. The user will provide a target object they believe they are holding. Your job is to rigorously check visual details to confirm identity and condition against that target.

Output Format: *Return a JSON object with exactly these keys:*

- *match_confidence*: “High”, “Medium”, “Low”, or “None”.
- *visual_evidence*: List of observed features (Brand name, flavor/variant, text, color).
- *condition_report*: Objective notes on dents, tears, or open seals.

By combining these outputs via the summarizer, ShelfScout provides both high-level confirmation (“Yes, this is a bag of apples”) and decision-relevant details (“price: \$3.99, expiry: 15th January”) in a single interaction, addressing BVI shoppers’ need for comprehensive product verification documented by [Lee et al. \[2021\]](#).

Chat Fallback

For *intent=chat*, ShelfScout bypasses all task-specific pipelines and directly invokes the VLM with a generic conversational prompt:

You are a precise vision assistant. Your goal is to answer the user’s question based strictly on visual evidence from the image.

Output Format: *Return a JSON object with exactly these keys:*

- *visual_reasoning*: A short internal thought checking if the user’s question can be answered by the image. If not, state why.
- *visual_data*: A detailed, objective description of the visual elements relevant to the question.

This provides a fallback mechanism for queries outside the intended scope of ShelfScout. If users ask questions unrelated to grocery shopping (e.g., “What is the weather like today?”) or when noisy environments prevent clear query parsing, the system routes these queries to the chat fallback,

without forcing misclassified inputs into inappropriate SLIV workflows.

5.1.4 Summarizer

Response generation is regulated through a dual-stage summarization strategy. Initial outputs to the existing summarizer that exceed a pre-defined length threshold (300 characters)

are passed to a secondary OpenAI’s GPT OSS Groq API call with a stricter length constraint, ensuring concise, structured JSON responses while preserving essential information.

5.1.5 Fallbacks

Several additional system-level optimizations were incorporated to improve robustness and responsiveness under real-world constraints. To mitigate rate limits encountered with Groq APIs, multiple accounts are used in a cascading manner, allowing requests to be dynamically routed when one account exceeds its quota. For object detection, while the locally hosted Qwen2.5-VL model running on an NVIDIA RTX 3090 GPU serves as the primary object detector, due to occasional latency spikes and timeouts under heavy load, a [Fireworks](#)-hosted Qwen2.5-VL model is used as a fallback to maintain consistent performance.

5.2 System Evaluation

To evaluate the updated architecture of ShelfScout, a series of tests were conducted by the researchers at [IGA Marche Verreault et Normandeau](#), a grocery store located in Downtown Montreal, Quebec, Canada. This location was selected based on three criteria:

1. Institutional Access: Store management permitted in-store experimentation involving simulated grocery tasks using AI-based visual assistance systems.
2. Environmental Representativeness: The store follows standard supermarket layouts with wide aisles, uniform lighting and conventional shelf arrangements that generalize to other chain grocery stores.
3. Logistical Feasibility: Proximity to McGill University enabled rapid iteration. Researchers could test each architectural change in the grocery environment without travel overhead, allowing for quicker design revisions.

Two aisles were chosen as controlled test environments: the produce aisle³ and the cereal

³The produce aisle was approximately 5 meters in length, consisting of fruit and vegetable displays with minimal labeling.

aisle within the dry goods section.⁴ Both aisles were used in their natural state, with no modifications to product placement or signage, to preserve ecological validity.

The end-to-end SLIV pipeline was then evaluated through a complete task sequence. Starting in the produce aisle, the system was queried for a scene description, followed by a navigation request to locate “Kellogg’s Krave” cereal. The system first guided the user from the produce section to the cereal aisle using far-field navigation and then transitioned to near-field guidance to support object-level localization and grasping. Following this, the system was queried for object verification (e.g., “Is this the Krave cereal?”), demonstrating its ability to provide product-specific details such as flavor variant and price information.

To evaluate the usability of the system’s feedback in helping a user accomplish the intended task, every one out of three interactions was conducted while the user was blindfolded, relying entirely on system-generated instructions for navigation and grasping. Across all of these trials, the system was able to complete the end-to-end grocery assistance pipeline, but performance was not consistently stable. While the system did not exhibit hallucinated product attributes during these evaluations, navigation performance was sensitive to environmental and system-level constraints. RTAB-Map produced incorrect navigation instructions in approximately three out of ten cases, which led the user off-course. These failures could not always be recovered within the same session, requiring termination of the trial. Also, the store’s basement location experienced network instability, with bandwidth fluctuating between 1-10 Mbps, which introduced latency in the RTAB-Map navigation loop. The combination of navigation errors and increased latency limited the feasibility of safely conducting a full user study with BVI participants.

These challenges are actively being addressed in ongoing work. A dedicated ARKit-only navigation pipeline discussed in Section 6.2, which would replace RTAB-Map as the primary far-field guidance method, is currently under development in the lab and is expected to reduce the navigation errors observed during these trials. Efforts are also underway to mitigate the network latency introduced by the store’s basement location, including the use of dedicated mobile data plans to ensure stable connectivity during future trials. These ongoing efforts are intended to support a full in-store user study with BVI participants in the near future.

⁴The dry goods aisle was also approximately 5 meters in length, featuring densely packed shelves of cereals in branded packaging requiring fine-grained visual discrimination.

Chapter 6

Conclusion

6.1 Discussion and Conclusion

This thesis set out to answer two questions: how assistive AI can be designed to mitigate hallucinations or convey uncertainty in complex environments such as grocery stores (RQ1), and how such systems can sustain continuous guidance without requiring users to repeatedly halt and re-initiate an interaction (RQ2). To address these questions, this work first proposed the SLIV Framework in Section 3.2. We hypothesized that a task-based decomposition of the grocery shopping activity, including scene description, object description, guidance and object validation, would not only optimize backend query routing for grounded responses, but also better reflect the practical usability needs of a BVI user navigating a store with AI assistance [He et al., 2024, Chang et al., 2025].

To validate this framework, a preliminary version of the ShelfScout system was developed and its responses were compared against three commercial baselines (Be My AI, Meta Ray-Ban smart-glasses and Gemini Live). Through a remote user study with eight BVI participants, feedback was gathered on the practical requirements for a suitable AI grocery shopping assistant, highlighted in Section 4.4. This feedback directly drove the iteration of ShelfScout’s final system design. To mitigate hallucinations (RQ1), the system was designed to handle user queries through intent and object presence checks and generate responses via structured JSON that forced explicit model reasoning before output. To address the need for continuous guidance (RQ2), the task-level decomposition was refined to support automatic hand-overs. Object description and validation were combined to eliminate redundant interactions, and far-field navigation was designed to automatically trigger near-field hand-level

guidance upon completion. This iterated system was structured to fulfill the three design goals highlighted in Section 3.1: a modular architecture, grounded responses and sequential guidance. The in-store evaluation described in Section 5.2 demonstrated that the system did not exhibit hallucinated product attributes during these trials, providing preliminary evidence that intent classification, presence verification and structured JSON reasoning can reduce hallucination frequency, thereby attempting to address RQ1. In addition, this architecture could complete the end-to-end SLIV pipeline, including transitions between far-field and near-field guidance, in several trials, thus attempting to address the continuous mobility issue identified in RQ2. However, it is important to acknowledge that this final design could not be validated through a real-world user study with BVI participants due to unreliable navigation responses and network connectivity issues causing unacceptable response latency. Hence, while this research provides insights on how to design assistive AI for continuous activities and mitigate hallucinations, it does not establish ShelfScout as a fully validated and deployment-ready assistive system for BVI grocery shopping.

6.2 Ongoing Developments

Motivated by the failure modes observed during the in-store evaluation described in Section 5.2, an ARKit-only pipeline for both near-field and far-field guidance is currently underway. RTAB-Map relying on a server-side API call for both localization and step-by-step instruction generation meant that every navigation update depended on a stable network connection between the smartphone and the backend. In the basement-level store used for testing, fluctuating connectivity introduced latency in this loop and, in some instances, contributed to incorrect or even outdated navigation instructions that could not always be rectified within the same session. Since far-field guidance is meant to operate continuously while the user is walking, even brief interruptions and delays in this feedback loop translate directly into a user being misdirected, a risk that is unacceptable for a system intended to support unsupervised, real-world use. A dedicated ARKit-only pipeline will address this failure mode. ARKit runs natively on-device, meaning localization and pose tracking would no longer depend on a round-trip call to an external mapping service, removing one of the bottlenecks that caused failures during testing. This would also bring far-field navigation onto the same underlying localization framework already used for near-field guidance described in

Section 5.1, rather than requiring the handoff between RTAB-Map and ARKit for each kind of guidance as implemented in the current architecture. Consolidating both guidance phases onto a single, local pipeline is expected to reduce latency and ensure real-time instruction updates.

Exploration of the ARKit-based solution constitutes ongoing research activities of other lab members and is thus outside the scope of this thesis.

6.3 Future Work

While the research presented in this thesis establishes a working architecture for mitigating hallucinations and providing continuous guidance, the system’s practical deployment limitations and user feedback point to three additional areas for future work.

First, the primary barrier to a real-world user study was infrastructural instability. Therefore, the immediate next step for this research is to test ShelfScout’s evolved design in a physical grocery store where stable, high-speed network connectivity is guaranteed. A stable connection could also improve RTAB-Map’s real-time performance and its capacity for continuous error recovery during active navigation tasks. Securing this stable environment would allow ShelfScout to be validated by BVI participants in a real-world setting.

In addition, because relying on RTAB-Map proved fragile in a physical environment, the dedicated ARKit-only navigation pipeline described in Section 6.2 will be evaluated in-store in the same conditions. Testing this as a complete substitute for RTAB-Map could provide a more lightweight and reliable guidance solution.

Finally, from a system design perspective, future iterations of ShelfScout could integrate customized user profiles. Feedback from the preliminary user study made it clear that what makes an assistive AI “suitable” is participant-dependent rather than universal. For example, the needs of a low-vision participant differ from those of a totally blind participant. Also, users have different thresholds for verbosity.¹ By incorporating user profiles, the backend could dynamically adjust its routing logic and response generation, allowing participants to shape the system to their specific accessibility needs and preferences.

¹Some users required highly descriptive responses, while others preferred concise and action-oriented instructions.

6.4 Broader Implications

Beyond the technical modifications outlined in Section 6.3, this thesis surfaces broader questions about the role of AI in assistive contexts that extend past the scope of ShelfScout itself.

6.4.1 Hallucinations

Although ShelfScout’s intent classification, presence checks and structured reasoning appeared effective during internal testing in mitigating hallucinated product attributes across the evaluated trials, this does not constitute a guarantee against hallucination in future use. The underlying VLMs remain unpredictable: a model that behaves reliably across dozens of trials may still hallucinate on the next query, and no architectural safeguard proposed in this thesis, or elsewhere in the literature, can rule this out entirely with existing technology. Current certification and safety standards for AI systems are built around aggregate accuracy benchmarks, which do not translate to a use case where a single incorrect response, however rare, can carry a physical safety cost. ShelfScout’s design reduces the likelihood of failure but does not eliminate it.

6.4.2 Accountability

This unpredictability points to a related and unaddressed question: who is responsible when an assistive AI system does hallucinate? ShelfScout, like Be My AI and Meta Ray-Ban smart-glasses, depends on VLMs hosted and maintained by third parties, in this case, via the Groq API, whose weights, availability and behavior are outside the researcher’s control. Section 5 noted that Meta’s Llama-4 Maverick was deprecated from Groq’s free tier partway through this project, forcing a switch to Meta’s Llama-4 Scout, without any change to ShelfScout’s own architecture. If a model update, deprecation or distillation introduces a new failure mode, an assistive AI system built on top of it inherits that risk without being aware of it. This raises an accountability gap: when a VLM hallucinates and misleads a BVI user into an unsafe action, existing legal and institutional frameworks offer little clarity on where responsibility lies between the model provider, the application developer and the user who acted on the system’s guidance. As assistive AI systems increasingly depend on

externally hosted, frequently updated VLMs rather than a single, static model, this absence of accountability structures becomes a more pressing gap.

6.4.3 Dependence vs. Autonomy

Finally, there is a trade-off between optimizing for seamless guidance and user agency: continuous, automatic hand-offs between tasks—the core contribution of the SLIV framework—reduce interaction overhead, but they also shift decision-making away from the user. A guide dog allows its handler to set the pace and choose a route, while ShelfScout, by design, determines when to transition from far-field guidance to near-field hand-level reaching guidance and then to validation. This thesis treated seamlessness as an improvement, motivated directly by user feedback requesting fewer, less repetitive interactions. However, it is worth asking whether an assistive system optimized for continuity risks diminishing the user’s sense of control over their own shopping experience, even as it reduces friction. Future evaluations of systems like ShelfScout may need to measure not only task completion and response quality, but also the degree to which users feel they remain in control of the interaction.

Bibliography

- Abdulahdy Abas Abdullah, Arkaitz Zubiaga, Seyedali Mirjalili, Amir H Gandomi, Fatemeh Daneshfar, Mohammadsadra Amini, Alan Salam Mohammed, and Hadi Veisi. Evolution of Meta’s Llama Models and Parameter-Efficient Fine-Tuning of Large Language Models: A Survey. *arXiv preprint arXiv:2510.12178*, 2025.
- Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b Model Card. *arXiv preprint arXiv:2508.10925*, 2025.
- Shivendra Agrawal and Bradley Hayes. GIST: Multimodal Knowledge Extraction and Spatial Grounding via Intelligent Semantic Topology. *arXiv preprint arXiv:2604.15495*, 2026.
- Groq AI. Llama 4 Maverick 17B 128E, 2025a. URL <https://console.groq.com/docs/model/meta-llama/llama-4-maverick-17b-128e-instruct>.
- Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation, 2025b. URL <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>.
- Fernando Alonso, José Fuertes, Ángel González Martínez, and Loïc Martínez Normand. User-Interface Modelling for Blind Users. volume 5105, pages 789–796, July 2008. ISBN 978-3-540-70539-0. doi: 10.1007/978-3-540-70540-6_117.
- Ramsha Azam and Mahfar Khan. Effectiveness Of Google Lookout App On Functional Task Performance In Low Vision Patients With Macular Disease. *Ophthalmology Pakistan*, 15: 105–111, December 2025. doi: 10.62276/OphthalmolPak.15.04.219.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin

- Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibor Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- Michela Balconi, Laura Angioletti, and Carlotta Acconito. “Choose with your eyes closed” instore shopping experience and spatial maps in visually impaired and sighted persons. *Univers. Access Inf. Soc.*, 24(1):853–865, April 2024. ISSN 1615-5289. doi: 10.1007/s10209-024-01115-0. URL <https://doi.org/10.1007/s10209-024-01115-0>.
- Marziyeh Bamdad, Davide Scaramuzza, and Alireza Darvishy. SLAM for Visually Impaired People: A Survey. *IEEE Access*, 2024.
- Roni Barak Ventura, Giles Hamilton-Fletcher, and John-Ross Rizzo. From abandonment to adoption: advancing assistive technologies for blindness and low vision in the AI era. *Frontiers in Digital Health*, Volume 7, 2025. ISSN 2673-253X. doi: 10.3389/fdgth.2025.1719746. URL <https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2025.1719746>.
- Be My Eyes. Introducing Be My AI (formerly Virtual Volunteer) for People who are Blind or Have Low Vision, Powered by OpenAI’s GPT-4. Company Blog, 2023. URL <https://www.bemyeyes.com/blog/introducing-be-my-eyes-virtual-volunteer>.
- Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, and Tom Yeh. VizWiz: Nearly Real-Time Answers to Visual Questions. In *Proceedings of the International Cross Disciplinary Conference on Web Accessibility (W4A)*, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450300452. doi: 10.1145/1805986.1806020. URL <https://doi.org/10.1145/1805986.1806020>.
- John M. Carroll, Michelle McManus, Sooyeon Lee, Peter A. Zientara, and Vijaykrishnan Narayanan. The Third Eye: A Shopping Assistant for the Visually Impaired. CHI EA, page 465, New York, NY, USA, 2017. Association for Computing Machinery. ISBN

9781450346566. doi: 10.1145/3027063.3049786. URL <https://doi.org/10.1145/3027063.3049786>.

Ruei-Che Chang, Rosiana Natalie, Wenqian Xu, Jovan Zheng Feng Yap, and Anhong Guo. 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, ASSETS*, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400706769. doi: 10.1145/3663547.3746319. URL <https://doi.org/10.1145/3663547.3746319>.

Anh-Hoang Dang et al. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence*, 8:1622292, September 2025. doi: 10.3389/frai.2025.1622292. URL <https://doi.org/10.3389/frai.2025.1622292>.

Kristina Deryagina, Claudia Müller, and Larissa Jesus. Design of a grocery shopping navigation support system for visually impaired individuals. *Mensch und Computer - Workshopband*, 2023.

Pei Du and Nirupama Bulusu. An Automated AR-Based Annotation Tool for Indoor Navigation for Visually Impaired People. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS*, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383066. doi: 10.1145/3441852.3476561. URL <https://doi.org/10.1145/3441852.3476561>.

Sandra Fernando, Chiemela Ndukwe, Bal Virdee, and Ramzi Djemai. Image Recognition Tools for Blind and Visually Impaired Users: An Emphasis on the Design Considerations. *ACM Trans. Access. Comput.*, 18(1), January 2025. ISSN 1936-7228. doi: 10.1145/3702208. URL <https://doi.org/10.1145/3702208>.

Bhanuka Gamage, Thanh-Toan Do, Nicholas Price, Arthur Lowery, and Kim Marriott. What do Blind and Low-Vision People Really Want from Assistive Smart Devices? Comparison of the Literature with a Focus Study. pages 1–21, October 2023. doi: 10.1145/3597638.3608955.

- Bhanuka Gamage, Leona Holloway, Nicola McDowell, Thanh-Toan Do, Nicholas Price, Arthur Lowery, and Kim Marriott. Vision-Based Assistive Technologies for People with Cerebral Visual Impairment: A Review and Focus Study. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–20, 2024.
- Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. VizWiz Grand Challenge: Answering Visual Questions from Blind People. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018.
- Tesfay Gidey Hailu, Xiansheng Guo, and Haonan Si. Indoor Positioning Systems as Critical Infrastructure: An Assessment for Enhanced Location-Based Services. *Sensors*, 25(16), 2025. ISSN 1424-8220. doi: 10.3390/s25164914. URL <https://www.mdpi.com/1424-8220/25/16/4914>.
- Yu Hao, Junchi Feng, John-Ross Rizzo, Yao Wang, and Yi Fang. Detect and Approach: Close-Range Navigation Support for People with Blindness and Low Vision. In *European Conference on Computer Vision*, pages 607–622. Springer, 2022.
- Junxian He, Shrinivas Pundlik, and Gang Luo. Can ChatGPT assist visually impaired people with micro-navigation? *arXiv preprint arXiv:2408.08321*, 2024.
- Pilar Herrero-Martín and Álvaro García-Ballester. Designing Human-Centred Adaptive AI Navigation for Blind and Visually Impaired Individuals: A Cognitive Load-Aware Framework for Accessible Urban Mobility. *MDPI AI*, 7(6), 2026. ISSN 2673-2688. doi: 10.3390/ai7060206. URL <https://www.mdpi.com/2673-2688/7/6/206>.
- hexgrad. Kokoro TTS: An open-weight TTS model with 82 million parameters. <https://huggingface.co/spaces/hexgrad/Kokoro-TTS>, 2025. Hugging Face Spaces. Accessed: September 11.
- Jie Hu, Changyuan Yan, Yu Zheng, Ziqian Wang, and Jiaming Zhang. Position: Assistive Agents Need Accessibility Alignment. *arXiv preprint arXiv:2605.13579*, 2026.

- Shaun K. Kane, Jeffrey P. Bigham, and Jacob O. Wobbrock. Slide Rule: Making Mobile Touch Screens Accessible to Blind People Using Multi-Touch Interaction Techniques. In *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets, page 73–80, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781595939760. doi: 10.1145/1414471.1414487. URL <https://doi.org/10.1145/1414471.1414487>.
- Shaun K. Kane, Jacob O. Wobbrock, and Richard E. Ladner. Usable Gestures for Blind People: Understanding Preference and Performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI, page 413–422, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450302289. doi: 10.1145/1978942.1979001. URL <https://doi.org/10.1145/1978942.1979001>.
- Doaa Khattab, Julie Buelow, and Donna Saccatelli. Understanding the Barriers: Grocery Stores and Visually-Impaired Shoppers. *Journal of Accessibility and Design for All*, 5: 157–173, November 2015. doi: 10.17411/jaccess.v5i2.105.
- Hyung Kim. User experience of assistive apps among people with visual impairment. *Technology and Disability*, 34:1–10, May 2022. doi: 10.3233/TAD-220377.
- Vladimir Kulyukin and Aliasgar Kutiyanaawala. Accessible Shopping Systems for Blind and Visually Impaired Individuals: Design Requirements and the State of the Art. *The Open Rehabilitation Journal*, 3:158–168, January 2010. doi: 10.2174/1874943701003010158.
- Bineeth Kuriakose, Raju Shrestha, and Frode Sandnes. Tools and Technologies for Blind and Visually Impaired Navigation Support: A Review. *IETE Technical Review*, 39:1–16, September 2020. doi: 10.1080/02564602.2020.1819893.
- Mathieu Labbé and François Michaud. RTAB-Map as an Open-Source Lidar and Visual Simultaneous Localization and Mapping Library for Large-Scale and Long-Term Online Operation. *Journal of Field Robotics*, 36(2):416–446, 2019.
- Patrick E. Lanigan, Aaron M. Paulos, Andrew W. Williams, Dan Rossi, and Priya Narasimhan. Trinetra: Assistive Technologies for Grocery Shopping for the Blind. In *10th IEEE International Symposium on Wearable Computers*, pages 147–148, 2006. doi: 10.1109/ISWC.2006.286369.

- Alexandru Lavric, Cătălin Beguni, Eduard Zadobrischi, Alin-Mihai Căilean, and Sebastian-Andrei Avătămăniței. A Comprehensive Survey on Emerging Assistive Technologies for Visually Impaired Persons: Lighting the Path with Visible Light Communications and Artificial Intelligence Innovations. *MDPI Sensors*, 24(15), 2024. ISSN 1424-8220. doi: 10.3390/s24154834. URL <https://www.mdpi.com/1424-8220/24/15/4834>.
- Kyungyeon Lee, Sohyeon Park, and Uran Oh. 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*, CHI EA, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380959. doi: 10.1145/3411763.3451806. URL <https://doi.org/10.1145/3411763.3451806>.
- Vincent T Lee, Tanfer Alan, Sung Kim, Ecenur Ustun, Amr Suleiman, Ajit Krishna, Tim Balbekov, Armin Alaghi, and Richard Newcombe. Full System Architecture Modeling for Wearable Egocentric Contextual AI. *arXiv preprint arXiv:2512.16045*, 2025.
- Chengpu Liao, Xingpan Jin, and EunJeong Cheon. EasyGO: A Field Study of Grocery Store Navigation Application Design for the Visually Impaired. In *Companion Publication of the ACM Designing Interactive Systems Conference*, DIS Companion, page 214–218, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706325. doi: 10.1145/3656156.3663719. URL <https://doi.org/10.1145/3656156.3663719>.
- Huei-Yung Lin, Yu-Hsiang Fan, and Chin-Chen Chang. Multimodal Navigation System for Visually Impaired Users Using Environmental Perception and Vision-Language Models. *MDPI Sensors*, 26(10), 2026. ISSN 1424-8220. doi: 10.3390/s26103045. URL <https://www.mdpi.com/1424-8220/26/10/3045>.
- Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A Survey on Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2402.00253*, 2024.
- Chen-Lung Lu, Zi-Yan Liu, Jui-Te Huang, Ching-I Huang, Bo-Hui Wang, Yi Chen, Nien-Hsin Wu, Hsueh-Cheng Wang, Laura Giarré, and Pei-Yi Kuo. Assistive Navigation Using Deep Reinforcement Learning Guiding Robot With UWB/Voice Beacons and Semantic Feedbacks for Blind and Visually Impaired People. *Frontiers in Robotics and AI*, 8:654132,

2021. doi: 10.3389/frobt.2021.654132. URL <https://www.frontiersin.org/articles/10.3389/frobt.2021.654132/full>.

Yaxi Lu, Haolun Li, Xin Cong, Zhong Zhang, Yesai Wu, Yankai Lin, Zhiyuan Liu, Fangming Liu, and Maosong Sun. Learning to Generate Structured Output with Schema Reinforcement Learning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4905–4918, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.243. URL <https://aclanthology.org/2025.acl-long.243/>.

Kelly Mack, Emma McDonnell, Dhruv Jain, Lucy Lu Wang, Jon E. Froehlich, and Leah Findlater. What Do We Mean by “Accessibility Research”? A Literature Survey of Accessibility Papers in CHI and ASSETS from 1994 to 2019. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2021.

Luis Marquez-Carpintero, Francisco Gomez-Donoso, Zuria Bauer, Bessie Dominguez-Dager, Alvaro Belmonte-Baeza, Mónica Pina-Navarro, Francisco Morillas-Espejo, Felix Escalona, and Miguel Cazorla. AIDEN: Design and Pilot Study of an AI Assistant for the Visually Impaired. *IEEE Access*, 2026.

Michele McDonnall and Rachael Sessler Trinkowsky. Assistive Technology Innovations: Perceptions, Adoption, and Desires. *Assistive Technology Outcomes and Benefits*, 19:1–16, January 2025.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yungang Jiang. DeepStack: Deeply Stacking Visual Tokens is Surprisingly Simple and Effective for LMMs. *Advances in Neural Information Processing Systems*, 37:23464–23487, 2024.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 20286–20332, 2025.

- Prudhvi Naayini, Praveen Kumar Myakala, Chiranjeevi Bura, Anil Kumar Jonnalagadda, and Srikanth Kamatala. AI-Powered Assistive Technologies for Visual Impairment. *arXiv preprint arXiv:2503.15494*, 2025.
- NaviLens. Empowering blind and partly sighted people. Company Blog, 2022. URL <https://www.navilens.com/en/>.
- Sama Nikanfar, Aref Hebri, Harish Ram Nambiappan, Gaurav Nale, Mahfuza Siddiqua, Farnaz Farhanipad, and Fillia Makedon. A Survey on Assistive Technologies for Visually Impaired Individuals: Recent Innovations, Limitations, and Future Directions. In *Proceedings of the 18th ACM International Conference on Pervasive Technologies Related to Assistive Environments*, PETRA, page 429–434, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714023. doi: 10.1145/3733155.3734895. URL <https://doi.org/10.1145/3733155.3734895>.
- Yeonjeong Park. A Pedagogical Framework for Mobile Learning: Categorizing Educational Applications of Mobile Technologies into Four Types. *The International Review of Research in Open and Distributed Learning*, 12(2):78–102, February 2011. doi: 10.19173/irrodl.v12i2.791. URL <https://www.irrodl.org/index.php/irrodl/article/view/791>.
- Umang Patel, Gerardo Moreno, Abhilash Tayade, and Princely Jonas Lopes. Artificial Intelligence Testing: Exploring Microsoft’s SeeingAI Across Four Scenarios. May 2024.
- Alec Radford and Karthik Narasimhan. Improving Language Understanding by Generative Pre-Training. 2018. URL <https://api.semanticscholar.org/CorpusID:49313245>.
- Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.
- André Rodrigues, Hugo Nicolau, Kyle Montague, João Guerreiro, and Tiago Guerreiro. Open Challenges of Blind People Using Smartphones. *International Journal of Human–Computer Interaction*, 36(17):1605–1622, 2020.
- Ligao Ruan, Giles Hamilton-Fletcher, Mahya Beheshti, Todd E Hudson, Maurizio Porfiri, and John-Ross Rizzo. A Multimodal Assistive System for Product Localization and Re-

- trieval for People who are Blind or have Low Vision. *arXiv preprint arXiv:2601.12486*, 2026.
- Manaswi Saha, Alexander J. Fiannaca, Melanie Kneisel, Edward Cutrell, and Meredith Ringel Morris. Closing the Gap: Designing for the Last-Few-Meters Wayfinding Problem for People with Visual Impairments. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS, page 222–235, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366762. doi: 10.1145/3308561.3353776. URL <https://doi.org/10.1145/3308561.3353776>.
- Hajar Sakai, Yi-En Tseng, Mohammadsadegh Mikaeili, Joshua Bosire, and Franziska Jovin. From Staff Messages to Actionable Insights: A Multi-Stage LLM Classification Framework for Healthcare Analytics. *arXiv preprint arXiv:2509.05484*, 2025.
- Patrícia Schmidtová, Niyati Bafna, Seth Aycok, Gianluca Vico, Wiktor Kamzela, Katharina Hämmerl, and Vilém Zouhar. How Important is Perfect English for Machine Translation Prompts? *arXiv preprint arXiv:2507.09509*, 2025.
- Suraj Singh Senjam. Smartphones as Assistive Technology for Visual Impairment. *Nature: Eye (London, England)*, 35(8):2078–2080, 2021. doi: 10.1038/s41433-021-01499-w.
- Serge Liatko. Prompt Engineering Is Dead, and Context Engineering Is Already Obsolete: Why the Future Is Automated Workflow Architecture with LLMs, 2025. URL <https://community.openai.com/t/prompt-engineering-is-dead-and-context-engineering-is-already-obsolete-why-the-future-is-automated-workflow-architecture-with-llms/1314011>.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarczyk, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *arXiv preprint arXiv:1701.06538*, 2017.
- Kristen Shinohara and Jacob O. Wobbrock. In the Shadow of Misperception: Assistive Technology Use and Social Interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI, page 705–714, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450302289. doi: 10.1145/1978942.1979044. URL <https://doi.org/10.1145/1978942.1979044>.

- Connor Shorten, Charles Pierse, Thomas Benjamin Smith, Erika Cardenas, Akanksha Sharma, John Trengrove, and Bob van Luijt. StructuredRAG: JSON Response Formatting with Large Language Models. *arXiv preprint arXiv:2408.11061*, 2024.
- R. Smith. An Overview of the Tesseract OCR Engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 629–633, 2007. doi: 10.1109/ICDAR.2007.4376991.
- Kaiyuan Tang, Yiqi Li, Xiaoyu Han, Ziyi Rui, Xing Zhou, and Qiuyu Lin. SeeingGrocer: A Smart Shopping Suite to Assist Visually Impaired Customers in Locating and Understanding Products. In *Companion Publication of the ACM Designing Interactive Systems Conference*, DIS Companion, page 217–221, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714863. doi: 10.1145/3715668.3735599. URL <https://doi.org/10.1145/3715668.3735599>.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*, 2023.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30, 2017.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
- Samantha A Wilson, Paula Byrne, Sarah E Rodgers, and Michelle Maden. A Systematic Review of Smartphone and Tablet Use by Older Adults With and Without Cognitive Impairment. *Innovation in Aging*, 6(2):igac002, January 2022. ISSN 2399-5300. doi: 10.1093/geroni/igac002. URL <https://doi.org/10.1093/geroni/igac002>.

- Junbin Xiao, Nanxin Huang, Hao Qiu, Zhulin Tao, Xun Yang, Richang Hong, Meng Wang, and Angela Yao. Egoblind: Towards Egocentric Visual Assistance for the Blind. *Advances in Neural Information Processing Systems*, 38, 2026.
- Yanran Xu, Klaus-Peter Zauner, and Danesh Tarapore. Blind-Wayfarer: A Minimalist, Probing-Driven Framework for Resilient Navigation in Perception-Degraded Environments. *arXiv preprint arXiv:2503.07492*, 2025.
- Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V. *arXiv preprint arXiv:2310.11441*, 2023.
- Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025.
- Tiancheng Zhao, Peng Liu, Xuan He, Lu Zhang, and Kyusong Lee. Real-time Transformer-based Open-Vocabulary Detection with Efficient Fusion Head. *arXiv preprint arXiv:2403.06892*, 2024a.
- Yi Zhao, Yilin Zhang, Rong Xiang, Jing Li, and Hillming Li. VIALM: A Survey and Benchmark of Visually Impaired Assistance with Large Models. *arXiv preprint arXiv:2402.01735*, 2024b.

Appendix A

Appendix

A.1 Recruitment Form

Call for Participation You are invited to participate in a research study involving AI-powered visual assistants for blind and visually impaired individuals in the Department of Electrical and Computer Engineering at McGill University.

Who can participate? You must be:

- Between 18 and 60 years old
- Able to provide informed consent
- Have reasonable mobility
- Familiar with assistive technologies (e.g., screen readers, magnifiers, etc.) or smart-phones
- Comfortable using voice-based systems or voice assistants
- Have no additional cognitive impairments that may affect task comprehension or communication
- Fluent in the language used for the study (English or French)

What is involved? You will be asked to:

1. Complete a demographic screening and pre-test questionnaire

2. Join a video conference meeting on Zoom and listen to several audio files of responses generated by different AI visual assistants when prompted with questions related to scene interpretation, locating a desired item in the environment, being guided toward it, and validating the identity of the item or interacting with it
3. Provide feedback regarding the audio responses

Unless you are a member of the Shared Reality Lab, you will be compensated \$20/hour, pro-rated to the nearest 15-minute interval. Members of the Shared Reality Lab do not receive compensation for participation in lab-conducted studies. If you withdraw before completing the activity, you will be compensated for the time you participated.

Interested in Participating? Please contact us here:

Principal Investigator: Henrique Jongh Pinheiro de Oliveira

Contact: henrique.j.oliveira@mail.mcgill.ca

If you have any concerns about this call for participation or study, please contact the principal investigator, Prof. Jeremy Cooperstock at jer@cim.mcgill.ca or by telephone at 514-398-5992.

A.2 Consent Form

REB File # 25-06-021 Participant Consent Form Multimodal AI-Powered Assistive System for Blind and Visually Impaired Users

You are being invited to participate in a study run by Henrique Oliveira (henrique.j.oliveira@mail.mcgill.ca) under the supervision of Professor Jeremy Cooperstock of the Department of Electrical and Computer Engineering of McGill University.

Research Personnel: Student researchers - Henrique J. P. de Oliveira, Mansi Dhanania, Kasra Jabbari, Melody Wang Supervisor - Jeremy R. Cooperstock

Purpose of the Study: This study aims to evaluate a multimodal, AI-powered assistive system that supports blind and visually impaired (BVI) individuals in performing daily living activities, including in-store grocery shopping, more independently.

Study Procedures: This study will be conducted either in a controlled environment such as our research laboratory or in a grocery store (with management permission). During the study, you will be asked to:

- Complete a pre-test questionnaire about demographics and technology experience - Use our SLIV system on a smartphone we provide to complete tasks such as: - Asking the system to describe scenes around you - Requesting the system to locate specific objects - Following audio guidance to approach objects - Verifying item details through the system - Think aloud while performing these tasks, verbalizing your thoughts and actions - Complete a post-test questionnaire about usability - Participate in a semi-structured interview about your experience

The session will last approximately 60-90 minutes. You will interact with the prototype through a smartphone application, with audio communication delivered either via the smartphone's speakers, your own headphones/earbuds, or bone conduction headphones provided by us. Suitable breaks will be provided between trials to avoid fatigue, and you may request additional breaks at any time. A researcher will remain within arm's reach at all times to ensure your safety.

Participant Eligibility: This study is open to both sighted (for preliminary testing) and BVI individuals who are familiar with using smartphones in daily activities, have reasonable hearing and mobility, and experience navigating typical indoor environments such as grocery stores or university buildings. Participants must be comfortable communicating verbally, as the system operates through spoken commands and delivers auditory feedback.

Voluntary Participation: Participation in this study is voluntary. You may at any point choose to decline to answer a question or withdraw from the study. Withdrawing from the study will not affect your relationship with the research team or prevent participation in future studies. You may withdraw by contacting us with your request and we will attempt to remove your data. However, if your request to withdraw comes after publication of the results of our study or after the deletion of identifying data, including the keyfile linking your name to your participant ID, we may not be able to do so.

Potential Risks: While the study is designed to be low-risk, you may experience the following: - Unfamiliarity with technology: You may feel uncertain or self-conscious while using a new voice-based system. We will provide clear instructions and reassurance throughout the study. - Physical discomfort: Some participants may find it physically or cognitively demanding to hold, orient, and interact with the device for extended periods. Accommodations, such as support from a nearby researcher or resting the phone on a stable surface, will be provided if required. - Cognitive fatigue: You may experience fatigue or frustration due

to sustained voice interaction and following audio instructions during the 90-minute session. You are encouraged to take short breaks as needed. - Audio sensory concerns: The prototype relies on spoken instructions which may be challenging in noisy environments. The study will be conducted in a reasonably quiet setting with audio delivered at comfortable volumes, and you may request repetition of instructions. - Interview stress: Some participants may feel pressure when asked to evaluate the prototype. The interview will be conducted in a relaxed format, and your honest input—positive or negative—is valuable.

For your safety, a member of the research team will remain within arm's reach at all times during the study.

Potential Benefits: You will not directly benefit from participating in this research. However, we hope that through your participation, this research will contribute to the development of accessible technologies that enhance independence for blind and visually impaired individuals in everyday activities.

Compensation: Unless you are a member of the Shared Reality Lab, you will be compensated \$20/hour, pro-rated to the nearest 15-minute interval. Members of the Shared Reality Lab do not receive compensation for participation in lab-conducted studies. If you withdraw before completing the activity, you will be compensated for the time you participated.

Confidentiality: Once you give consent for this study, you will be assigned a unique identifier (e.g., a code such as P01 or a pseudonym). This code will be used to link to your data. All identifiable information, including the mapping that links your code to you, will be kept secure and confidential, and only the researchers involved in this study will be able to see it. In scholarly activities (e.g., publications, presentations), you will only be referred to by your identifier.

Data Collection: During this study, we will collect audio recordings of your verbal responses and interactions with the system, smartphone sensor data (including camera images), and logs of system performance and user requests. Paper files with identifiable information (e.g., consent forms, honorarium receipts) will be scanned and stored digitally; the original paper documents will be destroyed. All non-anonymized digital data, (including your consent form, and honorarium receipt), will be kept on a secure, encrypted file server at the Centre for Intelligent Machines or on McGill's instance of Microsoft OneDrive, only accessible by the researchers involved in this study. Optionally, an anonymized transcript will be made from this recording, with all identifying information removed or redacted as appropriate, for

use in publication or future collaboration.

If you agree to this study, be aware that de-identified data may be kept indefinitely. We are therefore asking below if you consent to allow your de-identified data to be shared with other researchers in the future.

Yes _____ No _____ You consent to be audio recorded, for which these recordings will be accessed only by members of the research team.

Yes _____ No _____ You consent for your de-identified data, including demographics, experimental data consisting of Interaction logs (voice queries issued, system responses, task completion timestamps), System logs (object detection results, images captured), Task Performance data (success/failure of task completion through system response, task completion time, number of prompts issued to get required system response), Observational notes (hallucinations/errors in system response, notes on user behavior, assistance required), Audio recordings (if consented), transcripts and answers to questionnaires, to be shared with other researchers, for related future research purposes

Yes _____ No _____ You consent to be contacted about future experiments.

If you agree to use Interac e-Transfer to receive your compensation, your provided phone number or email address will be sent to the Interac platform in order to complete the transaction. To comply with Canada Revenue Agency Rules, your receipt for compensation (if provided in cash) or copy of the e-Transfer transaction will be retained for a period of 7 years. Any other identifying data will be destroyed within one year of the study's completion, which we estimate to be [insert date by which you expect the study to be complete].

Yes _____ No _____ You consent to be compensated via Interac e-Transfer

Right to Consult: You have the right to consult your study file in order to verify the personal information gathered, and to have it corrected if necessary, by contacting Henrique Oliveira, henrique.j.oliveira@mail.mcgill.ca.

Dissemination of Results: We expect to share the results of this research in academic publications, presentations, and reports. For such publications or future use of the data, you will be identified only by a numerical ID assigned to you.

Funding: This research is funded by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant. The funding agency has no direct involvement in the

study design, data collection, analysis, or publication of results.

Should you have any questions about this study, you may contact the research supervisor, Prof. Jeremy Cooperstock at jer@cim.mcgill.ca or by telephone at 514-398-5992.

If you have any ethical concerns or complaints about your participation in this study, and want to speak with someone not on the research team, please contact the Research Ethics Board Office at 514-398-5410 or daniel.tesolin@mcgill.ca citing REB file number [25-06-021].

Blind participants may give oral consent. Sighted participants are asked to sign below if you have read the above information and consent to participate in this study. Agreeing to participate in this study does not waive any of your rights or release the researchers from their responsibilities. To ensure the study is being conducted properly, authorized individuals, such as a member of the Research Ethics Board, may have access to your information. Please keep a copy of this document for your own reference.

For information regarding McGill's Privacy Notice, visit: <https://www.mcgill.ca/annex-mcgill-privacy-notice#research>

Participant's Name: _____ Participant's Signature: _____ Date: _____

A.3 Introductory Script

Hello, and thank you for joining us today. Before we begin, I'd like to ask whether you consent to us recording this session's audio and video for research purposes. The recording helps us analyze system performance, user interaction and navigation behavior. Do we have your consent to record?

[Pause for verbal consent.]

Today, we are evaluating two AI-based grocery assistance systems designed to support blind and low-vision shoppers in real stores: ShelfScout and Be My AI. This study helps us understand how well each system supports typical grocery tasks when used directly in a real store environment.

You will interact with both systems, each in a different aisle of the store. We are using two aisles with matched start-end distances: Produce aisle and Cereal aisle.

For each system you test, we'll ask you to complete these steps:

- 1. **Get oriented:** Ask the system to describe the scene around you to get a feel for the layout.*
- 2. **Find the item:** We'll give you a target object (like carrots or cereal). Interact with the system naturally, ask for directions or descriptions, and find and pick up the item. Think of this as a "simulated shopping experience"; we want to see how the system helps you move through the aisle and find what you need. (Note: No need to go to the checkout!)*
- 3. **Verify the product:** Once you've found the item, use the system to double-check the details. Feel free to ask about the price, ingredients or even the condition of the box to make sure it's exactly what you're looking for, simulating factors that help you make a purchase decision.*

You are encouraged to speak naturally, ask questions however you want and interact conversationally with the systems. We only ask that at some point during the interaction with each system, all tasks are performed with the aid of the system.

Think aloud throughout the process. We want to understand what makes sense, what doesn't and how you make decisions along the way.

For your safety and to support the procedure, we will accompany you during navigation and manage system operation, as well as ensure your safety. Another researcher will take observation notes on usability, errors, strategies and user reactions.

Important Notes Before We Begin:

- Your participation is voluntary. You may stop or withdraw at any time.*
- You will be compensated for your time. We can send your compensation via Interac deposit to the same email address we used to contact you, unless you prefer another method.*
- Your data will be kept confidential and anonymized.*
- You may request a break whenever needed.*

- *Please let us know immediately if at any point you feel unsafe, uncomfortable, or confused during navigation.*

Before we begin, do you consent to participate in this study as described?

[Pause for verbal consent.]

Great! Do you have any questions before we start?

A.4 Pre-Test Questionnaire

7/4/26, 3:40 PM

Pre-Test Questionnaire

Pre-Test Questionnaire

Participant Details and Background

* Required

1. Participant ID *

Enter your answer

2. Age (in years) *

Enter your answer

3. Gender *

- ☐ Woman
- ☐ Man
- ☐ Non-binary
- ☐ Prefer not to say
- ☐ Other

4. Vision Impairment *

- ☐ Mild Vision Impairment
- ☐ Moderate Vision Impairment
- ☐ Severe Vision Impairment
- ☐ Total Blindness
- ☐ Prefer not to say
- ☐ Other

5. Primary Occupation *

Enter your answer

https://forms.office.com/Pages/DesignPageV2.aspx?origin=NeoPortalPage&subpage=design&id=cZYxzedSaEqvqfz4-J8J6lKEdEjx_gdFtvKe1TBTi5B... 1/4

6. At what age did you first experience vision impairment, or were first diagnosed with your condition? *

- ☐ Early childhood (<=12 years)
- ☐ Adolescence (13-17 years)
- ☐ Early adulthood (18-30 years)
- ☐ Adulthood (31-60 years)
- ☐ Later in life (>60 years)
- ☐ I was born with my condition
- ☐ Prefer not to say

7. Do you rely on human assistance for navigation / support? *

- ☐ Yes
- ☐ No
- ☐ Maybe
- ☐ Prefer not to say

8. Do you use AI-powered systems for vision assistance (systems where you can take a picture/screenshot and it generates answers)? *

- ☐ Yes
- ☐ No
- ☐ Maybe
- ☐ Prefer not to say

9. Elaborate / Comments

Enter your answer

10. What types of Visual Assistance Technologies (VAT) do you currently use / have experience using?
(Please select all that apply) (options only include those relevant to our study and hence may not include technologies that the user uses in specific) *

- ☐ Be My Eyes
- ☐ Be My AI
- ☐ Seeing AI
- ☐ Aira
- ☐ Voice Over (iOS)
- ☐ Talk Back (Android)
- ☐ SeeForMe
- ☐ VoiceVista
- ☐ Lazarillo
- ☐ Meta Glasses
- ☐ No experience with VAT
- ☐ Prefer not to say
- ☐ Other

11. How often do you use AI-powered systems for vision assistance? *

- ☐ Very Often / Everyday
- ☐ Often / Few times a week
- ☐ Sometimes / Once a month
- ☐ Rarely / I have tried it a few times
- ☐ Never used AI for vision assistance
- ☐ Prefer not to say

12. How do the following statements align with your experience? *

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	N/A (Never Used)	Prefer not to say
I prefer human-assisted systems for navigation or support	☐	☐	☐	☐	☐	☐	☐
I prefer AI-powered systems for vision assistance	☐	☐	☐	☐	☐	☐	☐

13. Are there any specific challenges you face while using visual assistance technologies? Elaborate. *

Enter your answer

14. Is there any additional information about your vision impairment or technology usage that you believe might be useful for this research? *

Enter your answer

15. Any other insights / general details you would like to share? *

Enter your answer

This content is neither created nor endorsed by Microsoft. The data you submit will be sent to the form owner.



A.5 Post-Test Questionnaire

7/4/26, 3:40 PM

Post-Test Questionnaire

Post-Test Questionnaire

Participant Feedback

* Required

1. Participant ID *

Enter your answer

2. System (for result analysis) *

- ☐ System-A (Be My AI)
- ☐ System-B (Meta Ray Ban glasses)
- ☐ System-C (Gemini Live)
- ☐ System-D (ShelfScout)

3. Aisle (for result analysis) *

- ☐ Produce
- ☐ Dry Goods
- ☐ Bakery
- ☐ Cold Section

4. Which task had the most helpful response provided to help you complete it in a real world grocery store? *

- ☐ Scene Description
- ☐ Object Detection
- ☐ Guidance
- ☐ Validation
- ☐ None
- ☐ Prefer not to say

5. Why?

6. Did you feel confident that you could complete the task independently using this system? *

- ☐ Yes
- ☐ No
- ☐ Maybe
- ☐ Prefer not to say

7. Why?

8. On a scale of 1-completely unclear to 5-very clear, how understandable were the instructions? (3 - neutral/good) *



9. Was the guidance actionable (are you confident that you could act on it)? *

- ☐ Yes
- ☐ No
- ☐ Maybe / Partly
- ☐ Prefer not to say

10. Elaborate

11. On a scale of 1-very artificial to 5-very natural, how natural did the language and delivery of the response sound? (3 - neutral/good) *



12. Did anything in the audio response make you feel confused or uncertain? *

Enter your answer

13. On a scale of 1-very short to 5-very long, how did you feel about the length of the responses? (3 - neutral/good) *



14. On a scale of 1-lacking in detail to 5-too many unnecessary details, how did you feel about the content of the response? (3 - neutral/good) *



15. Were any details missing or unnecessary in the response? *

Enter your answer

16. Would you trust this system and use it for grocery shopping in real life? *

- ☐ Yes
- ☐ No
- ☐ Maybe
- ☐ Prefer not to say

17. Why?

Enter your answer

This content is neither created nor endorsed by Microsoft. The data you submit will be sent to the form owner.



A.6 Post-Study Semi-Structured Interview

7/4/26, 3:40 PM

Post-Study Semi-Structured Interview

Post-Study Semi-Structured Interview

Participant Feedback

1. Which system did you prefer overall (if any)? Why? [A/B/C/D]

Enter your answer

2. Was there a system that stood out as best for a specific task? Which and why?

Enter your answer

3. Overall, what had the most positive impact on you during this study?

Enter your answer

4. What had the most negative impact on you during this study?

Enter your answer

5. Was there a system's response structure that worked best for you for a specific task? Why?

Enter your answer

6. Any suggestions for making these tools more usable or helpful for blind/low vision shoppers?

Enter your answer

7. If you could design your own ideal system for independent grocery shopping, what would it do differently? What features would you add / remove? What attributes would make you feel most confident to use such a system?

Enter your answer

8. Are there situations where you'd prefer to receive human/AI assistance specifically, or where one feels more trust-worthy than the other? Why?

Enter your answer

9. Do you have any concerns about using AI-based assistance for everyday tasks?

Enter your answer

This content is neither created nor endorsed by Microsoft. The data you submit will be sent to the form owner.

