TOUVIGATION:

EMBODIED PROGRESSIVE GUIDANCE FOR BLIND AND LOW-VISION USERS

George X. Wang

Department of Educational Communication and Technology
EDCT-GE 2095 Capstone Thesis
Dr. Maaike Bouwmeester

New York University
xw3617@nyu.edu

April 29, 2026

Touvigation - Embodied Progressive Guidance for Blind and Low-Vision Users

GEORGE X. WANG, New York University, USA
JIAQIAN HU, Middlebury Institute of International Studies at Monterey, USA
SHAOYUE WEN, Imperial College London, The Hamlyn Centre, United Kingdom
JUNAN XIE, The Hong Kong University of Science and Technology (Guangzhou), China
JING QIAN*, Tongji University, College of Electronic and Information Engineering, China

CCS Concepts: • **Human-centered computing** → **Accessibility technologies**; *Interaction design*; • **Computing methodologies** → Computer vision; Natural language generation.

Additional Key Words and Phrases: blind and low-vision users, accessibility technologies, embodied interaction, multimodal large language models, egocentric wearable systems, assistive navigation, spatial guidance

1 Introduction

Recent advancements in Multimodal Large Language Models have substantially improved visual assistance for blind and low-vision (BLV) users (Be My Eyes, 2023, 2026, Penuela et al., 2026, Zhang and Lillianfeld, 2024). These tools can generate rich, natural-language descriptions of complex scenes for navigation that was previously difficult or impossible to obtain. Such capabilities have opened new doors for BLV users to improve everyday tasks, including navigation, object search, and situational awareness, supporting greater independence.

However, despite these promising developments, studies have found that current systems still remain misaligned with the practical needs of BLV users (Penuela et al., 2026, Zeraati et al., 2026). Prior work suggests that many LLM-powered tools are implicitly optimized for describing the world as perceived by sighted individuals, as the data trained were collected and labeled based on the sighted-centric assumption. This assumption introduces a series of challenges in real-world use for BLV users.

First, existing systems provide unstable and inconsistent spatial reference frames that was coherent for sighted users but not for BLV (Liu et al., 2024, Zeraati et al., 2026). Studies show that many tools alternately adopting scene-centric, object-centric, or camera-centric perspectives. They impose significant cognitive load on BLV users, who cannot rely on global visual alignment to reconcile these shifting frames rapidly. As a result, such descriptions often leads to unactionable guidance or even failures.

Second, these systems frequently rely on unembodied spatial metrics, such as metric distances (e.g., meters) or angular directions (e.g., degrees) (Ahmetovic et al., 2016, Messaoudi et al., 2022). For many BLV users, these abstractions lack intuitive meaning and are not readily translatable into bodily movement.

*Also with VIDA, New York University.

Authors' Contact Information: George X. Wang, xw3617@nyu.edu, New York University, New York, New York, USA; Jiaqian Hu, jiaqianh@middlebury.edu, Middlebury Institute of International Studies at Monterey, Monterey, California, USA; Shaoyue Wen, shaoyue.wen@imperial.ac.uk, Imperial College London, The Hamlyn Centre, London, United Kingdom; Junan Xie, jxie622@connect.hkust-gz.edu.cn, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, Guangdong, China; Jing Qian, jqian1590@tongji.edu.cn, Tongji University, College of Electronic and Information Engineering, Shanghai, China.

Without calibration to the user's body scale, reach, or locomotion patterns, these descriptions fail to support navigation or object interaction.

Consequently, despite their descriptive richness, current guidance systems often suffer from interaction inefficiencies, increasing user effort, time, and cognitive load, and in many cases leading to navigation breakdowns (Messaoudi et al., 2022, Zeraati et al., 2026). These limitations highlight a critical gap: existing approaches focus on improving what is described, while overlooking how descriptions translate into action (Dourish, 2001, Froese and Ziemke, 2009, Klemmer et al., 2006). This raises an important question: how can we design LLM-powered visual assistance systems that produce guidance directly grounded in users' embodied interaction with the environment?

To address this challenge, we introduce **Touvigation**, an egocentric wearable system that shifts the paradigm from passive snapshot descriptions to body-anchored, progressive guidance. Grounded in a formative study with ten BLV participants, Touvigation is designed around three core principles: (1) a **fixed body-centric reference anchor** that consistently treats the user as the spatial origin; (2) the use of **embodied metrics**, such as arm spans, clock-face directions, and step counts, which are directly mappable to physical actions; and (3) a **layered effective sensing envelope** based on user's navigation tools (hand, cane, and foot) to prioritize information based on actionable proximity.

By integrating these principles, Touvigation enables BLV-centered, proactive guidance powered by LLMs, effectively bridging the gap between AI perception and physical action. More broadly, our work contributes a shift in design perspective—from description-oriented systems to action-oriented, embodied guidance—offering a transferable framework for future assistive technologies.

2 Related Work

2.1 Visual Assistance Systems for BLV Users

Visual assistance systems for BLV users have evolved from automated computer vision pipelines to LLM-powered multimodal systems. Across this progression, the central goal has been to provide access to visual information.

Early systems largely relied on task-specific computer vision modules. Applications such as VizWiz (Bigham et al., 2010), Seeing AI (Microsoft Accessibility Blog, 2017), and Google Lookout (Google LLC, 2025) provide scene descriptions, text recognition, and object identification using on-device vision models, while wearable devices such as OrCam MyEye (OrCam, 2024) offer continuous audio-based access to visual information. More advanced systems integrate multimodal sensing; for example, Apple's Magnifier Detection Mode combines RGB cameras with LiDAR to detect people, doors, and objects (Apple, 2023). Despite these advances, such systems are typically constrained by predefined tasks and depend heavily on accurate camera framing, which remains a persistent challenge for BLV users. Consequently, outputs are often descriptive but not reliably actionable.

Some systems attempt to move beyond description toward guidance and interaction. For instance, Crosswatch (Coughlan and Shen, 2013) supports crosswalk detection in outdoor navigation, while NavCog3 (Sato et al., 2019) provides indoor navigation using beacon-based localization. Similarly, Soundscape (Microsoft Research, 2026) explores spatial audio to enhance environmental awareness rather than explicit turn-by-turn navigation. While these systems demonstrate the importance of closing the perception–action loop,

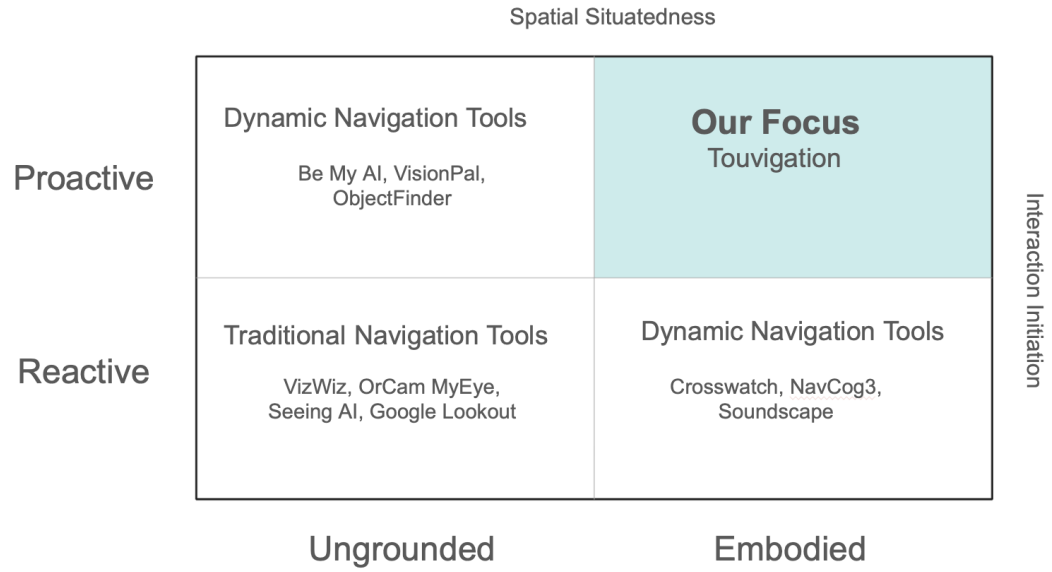

Fig. 1. A simplified design space of prior navigation support systems across two dimensions: reactive versus proactive guidance, and ungrounded versus embodied interaction. Traditional visual assistance tools largely remain reactive and ungrounded, while embodied navigation systems better support action but are often less proactive. Touvigation targets the proactive and embodied quadrant.

they often rely on specialized infrastructure or operate in constrained environments, limiting general applicability.

Recent advances in LLMs and Vision-Language Models have introduced a new paradigm of conversational visual assistance. Systems such as Be My Eyes (Be My Eyes, 2026), its Be My AI assistant (Be My Eyes, 2023), and generative image descriptions in Android TalkBack (Zhang and Lillianfeld, 2024) enable multi-turn interaction with visual content, allowing users to request task-specific information dynamically. Research prototypes further demonstrate this shift; for example, VisionPal (Penuela et al., 2026) shows that conversational querying improves user satisfaction and flexibility compared to static descriptions. However, LLM-powered systems introduce new challenges. Prior work identifies persistent issues such as hallucination and unreliable grounding in multimodal models (Azaria and Mitchell, 2023). Empirical studies with BLV users (Zeraati et al., 2026) highlight limitations including lack of camera guidance, difficulty interpreting spatial information, and inconsistent reference frames. Emerging systems such as ObjectFinder (Liu et al., 2024) attempt to integrate detection with LLM reasoning for object search and navigation, yet reliability in real-world scenarios remains inconsistent.

Overall, while both non-LLM and LLM-based systems have significantly improved access to visual information, they remain largely description-oriented. The challenge of translating perception into stable, embodied, and actionable guidance remains insufficiently addressed. This research gap directly leads to new approaches that explicitly integrate body-centered interaction and spatial grounding.

2.2 Spatial Guidance and Navigation Modalities

Spatial guidance and navigation for BLV users have been widely explored through multimodal assistive technologies that translate environmental information into accessible sensory cues. A dominant paradigm in prior work is sensory substitution, where auditory and haptic feedback are used to convey spatial information typically obtained through vision (Messaoudi et al., 2022). Such systems range from simple obstacle detection to complex navigation aids that integrate localization, path planning, and real-time feedback.

Audio-based guidance remains one of the most established approaches, particularly in mobile navigation applications. For instance, spatialized audio cues have been shown to support environmental awareness and waypoint navigation, enabling BLV users to explore surroundings more independently (Liu et al., 2022). However, reliance on auditory channels can interfere with users' ability to perceive critical environmental sounds, motivating research into alternative or complementary modalities (Bharadwaj et al., 2019).

Haptic guidance has emerged as a promising direction, offering continuous and unobtrusive directional feedback. Prior work demonstrates that wearable or cane-based vibrotactile systems can effectively communicate turn-by-turn instructions and obstacle proximity, particularly in noisy environments (Chaudary et al., 2023). More recent systems increasingly adopt multimodal feedback, combining audio, haptics, and voice interaction to improve robustness and usability across diverse contexts (Malaekah et al., 2026).

Despite these advances, prior work continues to face challenges in delivering actionable, context-aware guidance. Limitations include imprecise localization, cognitive overload from multimodal feedback, and insufficient support for last-meter navigation and dynamic environments (Messaoudi et al., 2022). Collectively, this body of work underscores the need for navigation systems that not only convey spatial information but also align with users' embodied practices and real-world mobility constraints.

2.3 Embodied and Actionable Interaction Design

Embodied and actionable interaction design for totally blind users centers on tight perception–action coupling, where system outputs directly support bodily action rather than passive interpretation (Dourish, 2001, Klemmer et al., 2006). Embodiment is grounded in physical and social practice, while actionability can be understood as providing timely feedforward, feedback, and recovery aligned with users' ongoing activities such as walking, touching, or typing (Ahmetovic et al., 2016, Froese and Ziemke, 2009). Across domains, the most effective systems are those that integrate with existing nonvisual skills and support closed-loop interaction. A primary strategy is skill transfer, where interfaces leverage established embodied routines. Braille-based mobile text entry systems such as Perkinput (Azenkot et al., 2012) and BrailleTouch (Southern et al., 2012) map multi-touch input to braille chords, enabling rapid learning and high performance through reuse of motor programs. Similarly, VR systems like Canetroller (Zhao et al., 2018) and subsequent cane-based controllers (Siu et al., 2020)replicate white-cane techniques, allowing users to explore virtual environments through familiar actions such as sweeping and tapping. These systems demonstrate that embodiment is most effective when grounded in practiced sensorimotor skills. A second pattern is actionable guidance as a control-loop problem, particularly in navigation. Systems such as NavCog (Ahmetovic et al., 2016) and NavCog3 (Sato et al., 2019) show that effective guidance requires continuous coupling between instruction and user movement, including support for error recovery and preview. Studies of rotation error

Table 1. Participant demographics from the formative interview study ($N = 8$). Vision impairment levels follow the Standard Chinese Vision Impairment Classification, where Level 1 indicates more severe vision loss and Level 4 indicates lower severity.

Participant	Gender	Age	Education Level	Blind School	Vision Impairment Level*	Onset of Impairment	Occupation	Prior AI Assistance Experience
P1	Male	30	PhD	Yes	1	Since age 4	Student	Yes
P2	Female	30	Undergraduate	No	1	Since 2021	N/A	Yes
P3	Female	27	Graduate	No	4	Since birth	Teacher	Yes
P4	Male	19	Undergraduate	No	4	Since age 9	Student	Yes
P5	Female	25	Undergraduate	No	1	Since 2023	N/A	Yes
P6	Male	23	Undergraduate	Yes	1	Since age 6	Massage Therapist	Yes
P7	Male	31	Vocational School	Yes	2	Since birth	Massage Therapist	Yes
P8	Male	24	Graduate	No	3	Since age 3	Student	Yes

further reveal that actionable interaction must account for human motor behavior, such as systematic overshoot during turns (Ahmetovic et al., 2018). Wearable and robotic systems, including Headlock (Fiannaca et al., 2014) and CaBot (Guerreiro et al., 2019), extend this by embedding guidance directly into bodily experience through haptic or physical feedback. A third theme is multimodal distribution of cognitive load in complex information access. Systems such as MAIDR (Seo et al., 2024) and Infosonics (Holloway et al., 2022b) combine braille, sonification, and text to support user-controlled exploration and verification, enabling users to actively interrogate data rather than passively consume descriptions (Seo et al., 2024; Muttler et al., 2022). Tactile displays further highlight embodied exploration strategies, where users position their hands to detect changes over time (Holloway et al., 2022a). Across prior work, most systems improve either dynamic access to visual information or embodied support for navigation, but rarely both at once. As a result, guidance often remains reactive, inconsistently grounded, or difficult to translate into immediate action. Touvigation builds on this gap by combining proactive assistance with body-centered reference frames, embodied metrics, and reachability-aware guidance, positioning it as an embodied and action-oriented alternative to existing approaches.

3 Formative Study

3.1 Participants

We conducted a formative interview study with eight blind or low-vision (BLV) participants (P1–P8). Participants were recruited through a social media post on RedNote and through direct outreach to blind users from the researchers' own contacts. The interviews were conducted remotely via Zoom or Tencent Meeting, depending on participants' preferences and accessibility needs. All participants agreed to audio recording for later transcription and analysis.

Participants ranged in age from 19 to 31 years old and represented diverse educational backgrounds, occupations, onset histories, and levels of vision impairment. Following the Standard Chinese Vision Impairment Classification, impairment levels ranged from Level 1 to Level 4, where Level 1 indicates more severe vision loss and Level 4 indicates lower severity. Three participants had experience attending blind schools, and all participants reported prior experience using AI-based visual assistance tools.

3.2 Data Collection

Data collection was qualitative in nature and centered on remote semi-structured interviews. All sessions were audio recorded with participants' consent, and later transcribed for analysis. The interviews focused on participants' existing strategies for visual assistance, their experiences with current AI-powered tools, and their expectations for future navigation support.

3.3 Procedure

Each interview consisted of three parts:

(1) **Background and AI tool usage:** Participants were asked about their backgrounds and prior experience with AI-based assistance tools, including which tools they used, how they used them, and the kinds of tasks they typically completed with them.

(2) **Think-aloud hotel scenario task:** Participants were asked to imagine that they had just checked into a hotel room and needed to complete two practical tasks: turning on the air conditioner and locating the trash can. They were asked to describe, step by step, how they would accomplish these tasks without AI tools and how they would do so with AI tools.

(3) **Open-ended design reflection:** Participants were invited to describe their ideal navigation tool if advanced AI capabilities were available, including what kinds of guidance, interaction styles, and environmental understanding would be most useful to them.

3.4 Challenges in the Current LLM-powered Tools

Based on our formative interviews, we summarized three core challenges that participants repeatedly encountered when using current VLM- and LLM-based visual assistance tools. Across participants' accounts, these systems were often effective at generating descriptions of the environment, but much less effective at supporting immediate, actionable navigation and interaction.

C1. Unstable reference anchors increase cognitive load. Participants described that current systems often switch unpredictably between scene-centered, object-centered, and body-centered reference frames. For example, a system might first describe an object relative to the overall room, then relative to another object, and then relative to the camera view. While these descriptions may be linguistically correct, they are difficult to operationalize because users cannot rapidly align multiple shifting reference frames through a single glance in the way sighted users can. As a result, participants reported having to mentally reconstruct the scene before acting, which substantially increased cognitive load and often made the guidance unusable in the moment.

C2. Unembodied metrics make instructions hard to execute. Participants also noted that many systems rely on abstract, vision-oriented spatial expressions such as "1–2 meters away" or "slightly to the left front." For users who are blind from birth or who have lived with long-term blindness, such measurements are often not intuitively meaningful. Even when the wording is technically precise, it does not readily map onto bodily action. Participants emphasized that descriptions become actionable only when they can be translated into directly perceivable movement units, such as steps, arm reach, or other body-scaled cues. Without that grounding, the system effectively fails at the point of execution.

C3. Granularity mismatch ignores users as embodied and augmented agents. A third challenge concerned the mismatch between the level of detail in system output and users' actual interaction capabilities. Existing systems often assume a uniform user model, but participants described navigation as an embodied and tool-mediated activity shaped by what they are currently using, such as a cane, a phone, or their hands. Whether one hand is occupied, whether the phone is being used for speech or camera input, or whether the cane is actively probing the environment all changes what kind of instruction is practical. However, current systems typically produce the same description regardless of these tool states. Participants therefore experienced a mismatch between informational granularity and actionable capability: some outputs were too coarse to support safe action, while others were too detailed to be useful in real time. In this sense, the description itself became a bottleneck in navigation rather than a form of assistance.

4 System

4.1 Design Process and Methodological Rationale

Table 2 summarizes the iterative research-through-design process, including the timeline, methods, and methodological rationale from Fall 2025 to Spring 2026. The process moved from broad exploration of AI-assisted BLV interaction toward a focused system contribution around embodied, actionable guidance.

4.2 Design Space and Rationale

To address the core challenges identified in our formative study, we developed **Touvigation** around three design principles. As shown in Figure 2, these principles frame visual assistance as a problem of supporting *embodied, proactive guidance* to help translate perception into instructions that are consistent, physically interpretable, and immediately actionable.

The rationale behind these principles is grounded in cognitive science and learning theory. Embodied cognition argues that thinking is shaped by sensorimotor capacities and bodily action, which supports instructions expressed in body-scaled units rather than abstract visual measurements. Cognitive load theory distinguishes task difficulty from extraneous load introduced by representation; unstable spatial frames and unnecessary unit conversion add extraneous load during real-time navigation. Situated action emphasizes that action unfolds in local context rather than from a complete plan alone, motivating guidance that updates according to the user's current posture, tools, and nearby affordances. Finally, distributed cognition treats cognition as distributed across people, artifacts, and environments, which is especially relevant for BLV mobility because hands, cane, feet, headset, and environment jointly support perception and action. Together, these perspectives explain why Touvigation prioritizes stable anchors, embodied metrics, and reachability-aware sensing rather than richer scene descriptions alone (Dourish, 2001, Hutchins, 1995, Suchman, 1987, Sweller, 1988, Wilson, 2002).

D1: Fixed Reference Anchor Through Body-Centricity

To address the unstable reference anchor problem (C1), we adopt a fixed body-centric frame in which the user is always treated as the spatial origin. The system only describes objects and directions relative to the user's current stance and forward-facing body orientation. The system avoids switching to object-centric statements such as "under the table" or camera-centric statements such as "left side of the image," and instead keeps all objects defined relative to the user's egocentric view. In particular, directional orientation

Table 2. Design process, timeline, and methodological rationale for Touvigation.

Stage	Timeline	Phase and Focus	Methods	Rationale for Methods	Key Outcomes
1. Opportunity Discovery	Sep. 1–Oct. 5, 2025	Early exploration of AI-assisted BLV interaction	Secondary research; feasibility analysis; early ideation	Appropriate for scoping emerging technologies, including LLMs, AR/VR, and multimodal sensing, and identifying constraints before defining a problem	Initial concepts (EyeQ/Guidion); hypotheses on gaze, head, and hand input
2. Research Planning	Oct. 6–Nov. 5, 2025	Structuring user-centered inquiry	Research planning; interview protocol design; stakeholder mapping	Ensures systematic investigation and avoids assumption-driven design	Defined research questions, target users, and study design (M2–M3)
3. Formative User Research	Nov. 6–Dec. 10, 2025	Understanding real-world practices and breakdowns	Semi-structured interviews; scenario-based elicitation; tool comparison	Captures contextual, lived experiences and reveals breakdowns in action-level tasks	Insights into limitations of current AI tools, including framing, distance, and usability issues
4. Synthesis and Insight Generation	Dec. 11–Dec. 20, 2025	Abstracting patterns across tasks	Thematic analysis; affinity diagramming	Enables generalization from diverse examples into core interaction challenges	Three challenges identified: unstable anchors, unembodied metrics, and granularity mismatch
5. Scope Narrowing and Reframing	Dec. 15, 2025–Jan. 20, 2026	Defining core contribution	Design critique; collaborative discussions; problem reframing	Prevents scope fragmentation and shifts from task-specific solutions to an interaction-level contribution	Reframed focus as embodied, actionable guidance, confirmed in Dec. 15 and Jan. 16 meetings
6. Literature Review and Positioning	Jan. 5–Feb. 5, 2026	Situating within prior work	Prior work review; design-space mapping; theoretical framing	Establishes novelty and connects the system to accessibility and HCI theory	Positioned system as proactive, embodied guidance
7. Design Principle Development	Jan. 20–Feb. 20, 2026	Translating insights into design rules	Research-through-design synthesis; body-coordinate modeling	Grounds design decisions in empirical findings and supports generalizable design claims	Three principles: body-centric reference, embodied metrics, and sensing envelope
8. Concept Development	Feb. 10–Mar. 5, 2026	Exploring interaction and use cases	Storyboarding; scenario design; interaction prototyping	Scenarios capture the temporal, action-based interaction needs of guidance systems	Defined interaction flows for navigation, object finding, and appliance use
9. Prototype and System Development	Mar. 1–Apr. 10, 2026	Implementing system concept	Hardware prototyping; modular architecture; prompt engineering	Validates feasibility and separates route-level navigation (Navigator) from near-field interaction (Toucher)	Functional prototype; wearable system; guidance grammar
10. Demonstration and Evaluation Planning	Apr. 1–Apr. 20, 2026	Communicating and preparing validation	Scenario-based demos; reflective analysis; evaluation design	Demonstrates dynamic embodied interaction and prepares for future empirical validation	Demo scenarios in kitchen and hotel contexts; evaluation framework confirmed in Apr. 3 meeting
11. Final Synthesis and Thesis Writing	Apr. 10–Apr. 25, 2026	Consolidation and documentation	Writing; diagramming; iterative refinement	Ensures clarity, contribution framing, and academic rigor	Final thesis, paper draft, and system articulation

is expressed using clock-face references anchored to the user's body, such as "at 3 o'clock" or "at 5 o'clock," rather than terms tied to the room or image. Regardless of where the user is, all instructions should be grounded in the same consistent anchor. By fixing the reference origin to the user's body, we aim to reduce mental remapping and improve the immediate executability of guidance. From a cognitive load perspective, the fixed anchor reduces extraneous load by removing the need to translate among room-centered, object-centered, and image-centered frames while the user is acting.

D2: Embodied Metrics for Actionable Guidance

To address the confusion caused by unembodied metrics (C2), we aim to translate spatial information into units that are directly meaningful through bodily action. Instead of using abstract visual measurements such as meters and degrees, the system uses body-scaled and movement-based expressions such as steps,

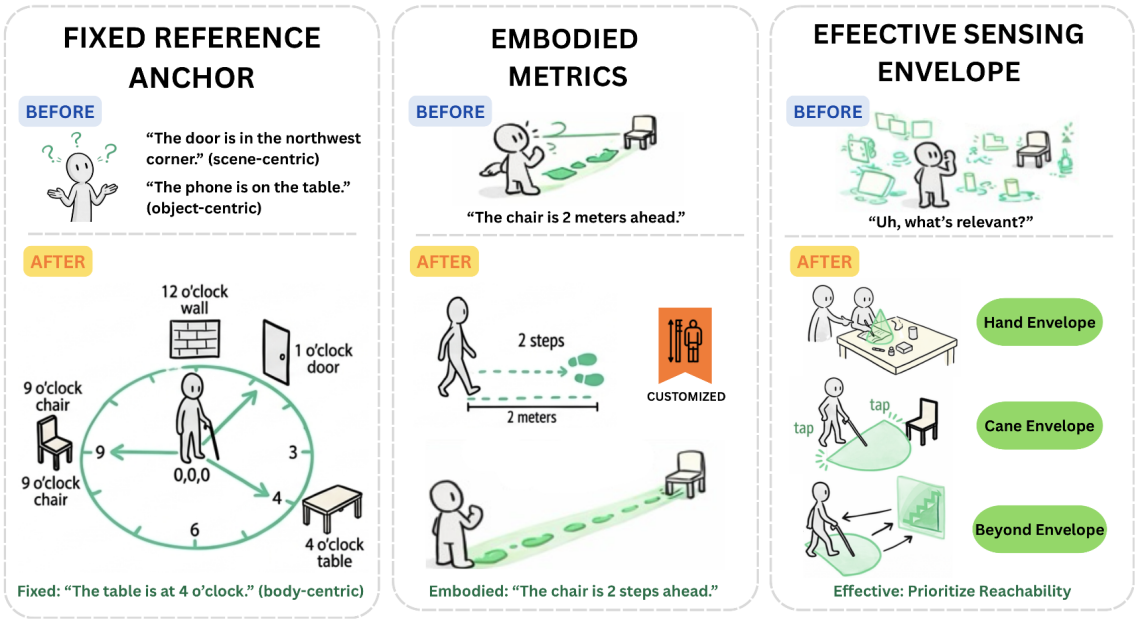

Fig. 2. Design space of navigation support systems

hand spans, cane sweeps, and body-height zones. The goal is to ensure verbal guidance can be converted into movement without requiring users to interpret abstract measurements. For example, distance can be expressed as "two steps" or "one cane sweep" rather than "1.5 meters," and size as "one hand span" rather than "20 centimeters wide." In this way, descriptions are phrased in terms that can be enacted or verified through the body itself. This design follows embodied cognition by treating the body not only as an output channel for instructions, but as the medium through which spatial information is perceived, learned, and verified.

D3: Effective Sensing Envelope for Reachability-Aware Guidance

To address the granularity mismatch (C3), we want to guide according to what the user can meaningfully sense and act on with their current body and tools. We conceptualize this as an *effective sensing envelope*: a layered interaction space shaped by the user's hands, cane, feet, and other available tools. Rather than producing a fixed-detail snapshot of the entire scene, the system prioritizes information based on reachability and action relevance. This principle reflects situated and distributed cognition: guidance is useful only when it fits the user's current activity context and the distributed system of body, mobility tools, wearable sensing, and environmental affordances.

This principle distinguishes between at least three layers: a hand envelope (roughly 0–70 cm) for fine manipulation and texture confirmation, a cane envelope (about 1–2 steps lookahead) for immediate ground-level obstacle detection, and a beyond-envelope preview layer for proactive mapping. To reduce cognitive load, only information that falls within the user's current effective envelope should be phrased as action-level instruction. Information outside the envelope should be presented at most as advance notice plus a tactile verification strategy, such as warning that a chair is ahead and suggesting confirmation with the next cane

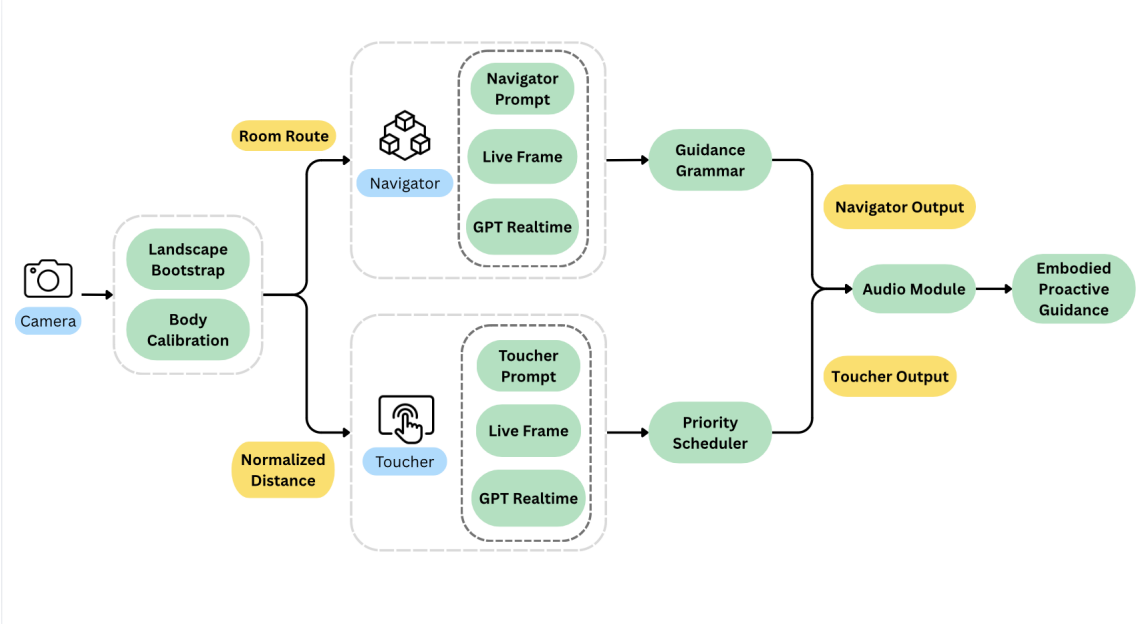

Fig. 3. System workflow of Touvigation (derived from logic model).

sweep before turning. The type of description should depend on envelope type: immediate, executable commands for reachable space, and preview-oriented guidance for non-reachable space. This allows the system to match detail to current action capacity, reducing overload while preserving environmental awareness.

4.3 Workflow

Figure 3 shows the runtime workflow of Touvigation. The system takes continuous camera input and user prompts, performs a one-time initialization step, and then routes processing into two parallel components: *Navigator* and *Toucher*. Each component combines live frames with a task-specific prompt and GPT Realtime, then sends its output to a shared audio module.

4.4 Initialization

Initialization prepares the shared context before walking begins. First, the landscape bootstrap module asks the user to perform a 360-degree scan so the vision pipeline can observe the surrounding space and build an initial room representation. We implement this stage in Python by sampling frames from the scan, stitching them into a panoramic view with OpenCV's Stitcher API (OpenCV, 2026), and applying Segment Anything (SAM) (Kirillov et al., 2023) to segment major objects in the scene. During the same setup stage, the system also asks for the user's current goal, such as the object they want to find or the destination they want to reach.

The route planner module then uses the stitched panorama, the segmented objects, and the user goal to generate a structured room-route JSON. This representation stores the room layout, detected objects, and a stepwise route plan. The room-route JSON is then passed to the Navigator as high-level route context.

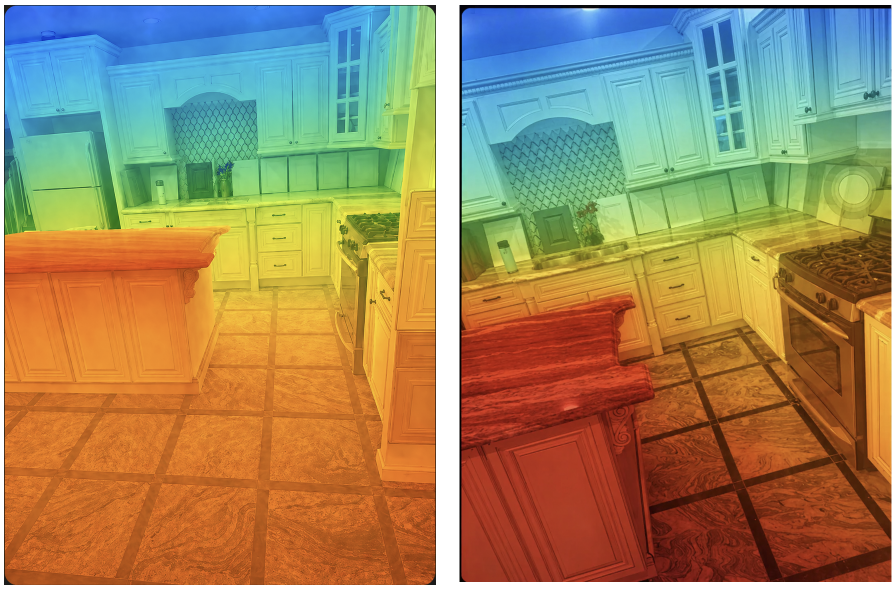

Fig. 4. Example Toucher visualization for near-range prioritization. As the user approaches the kitchen island and surrounding counters, nearby regions become more salient because their collision risk and immediate relevance increase.

In parallel, the body calibration module collects user-specific body metrics, such as height, hand length, arm length, foot size, and any mobility tools currently in use. We implement this stage with a calibration photo and body landmarks extracted with ML Kit Pose Detection (Google, 2026), combined with user-provided measurements when needed. These measurements are converted into a normalized-distance JSON that defines body-scaled units for later instruction generation. The normalized-distance JSON is then passed to the Toucher as calibration context.

4.5 Toucher

The Toucher module handles near-range, reachability-aware guidance. It samples the live camera stream at one frame per second (1 fps) and combines each frame with a Toucher prompt before sending it to GPT Realtime.

Toucher takes the live egocentric camera frame and uses a prompt-driven GPT Realtime pipeline to continuously assess nearby regions based on hazard, urgency, reachability, task relevance, and confidence. It then converts that ranked spatial information into adaptive feedback, warning the user about dangerous or important items and surfacing the target when it becomes visually clear and physically reachable. Figure 4 illustrates how nearby structures become more salient as the user moves closer.

To control when and what to say, Toucher applies a priority scheduler before forwarding output to audio. For each candidate guidance item i, the scheduler computes

$$P_i = \alpha\,\tilde{H}_i^{\,w_h} \;+\; \beta\,\tilde{U}_i^{\,w_u} \;+\; \gamma\,\tilde{R}_i^{\,w_r} \;+\; \delta\,\tilde{T}_i^{\,w_t} \;+\; \epsilon\,\tilde{C}_i^{\,w_c} \;+\; \lambda\,(\tilde{H}_i\tilde{U}_i\tilde{R}_i),$$

where $\tilde{H}_i$, $\tilde{U}_i$, $\tilde{R}_i$, $\tilde{T}_i$, and $\tilde{C}_i \in [0,1]$ are the normalized scores for hazard, urgency, reachability, task relevance, and confidence, respectively. The coefficients $\alpha, \beta, \gamma, \delta, \epsilon$ control the relative contribution of each term, w_h, w_u, w_r, w_t, w_c are importance exponents, and the interaction term $\lambda(\tilde{H}_i \tilde{U}_i \tilde{R}_i)$ boosts items that are simultaneously dangerous, urgent, and reachable.

Hazard is assigned from the detected object class and its collision severity. We use SAM object regions together with GPT-based scene parsing to mark obstacles such as walls, chairs, tables, and low-lying objects, then map each class to a normalized hazard score. Reachability is computed by comparing the estimated object distance against the normalized-distance JSON from body calibration. In implementation, per-frame object distance is estimated with Depth Anything V2 (Yang et al., 2024), and the resulting depth value is compared against the calibrated hand, cane, and foot ranges to determine whether an object falls within the current 1–2 step interaction window. Urgency is computed from time-to-contact or steps-to-contact, using the estimated object distance from Depth Anything V2 together with the user's recent walking speed or step length. Task relevance is computed by matching the current user goal against the detected object or region description. Confidence is computed from the agreement between visual detection confidence and the GPT output confidence. At each update cycle, the system normalizes all factor scores, computes P_i, ranks candidate items, selects the top item or top few items, and emits them as toucher output.

For example, the kitchen island receives a higher priority report as it becomes closer, because both urgency and collision risk increase as it moves into the user's immediate walking path.

The runtime loop is shown below:

```
while system_is_running:
  frame = sample_frame(fps=1)
  masks = SAM(frame)
  depth_map = DepthAnythingV2(frame)
  candidates = detect_items(masks, depth_map)
  for item in candidates:
    H = compute_hazard(item.class)
    U = compute_urgency(item.distance, user_speed, step_length)
    R = compute_reachability(item.distance, body_json)
    T = compute_task_relevance(item, user_goal)
    C = compute_confidence(item)
    item.priority = scheduler(H, U, R, T, C)
  output = select_top_items(candidates)
  send_to_audio(output)
```

4.6 Navigator

The Navigator branch handles route-level guidance. It takes the room-route JSON from initialization as persistent route context, then combines it with a Navigator prompt, a live frame sampled at one frame per second (1 fps), and GPT Realtime. The goal of this branch is to convert the planned route into short, structured navigation steps that can be updated online as the user moves.

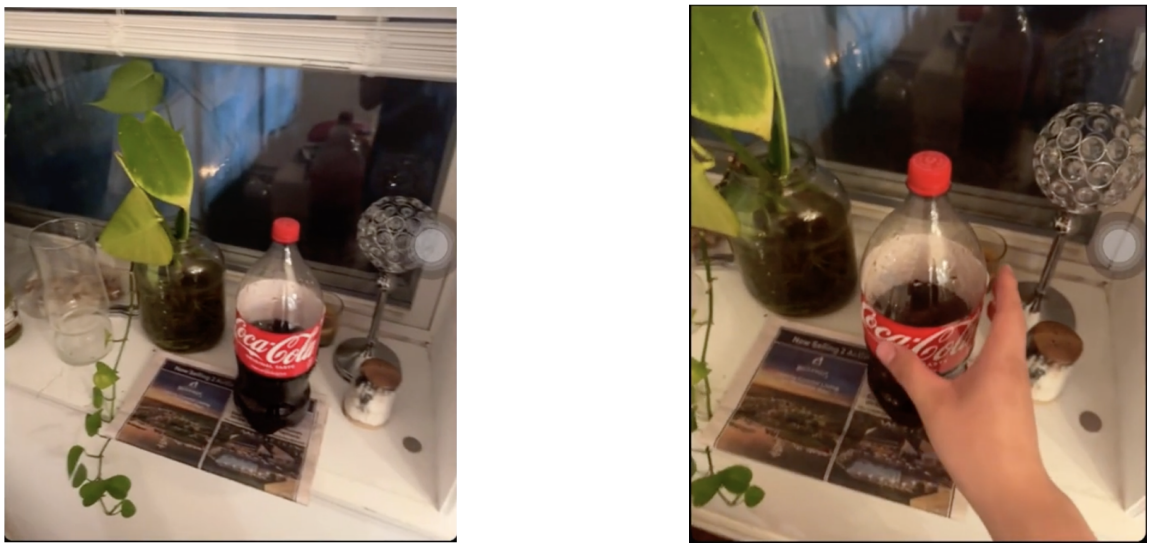

Fig. 5. Example Navigator sequence for object-directed guidance. The left frame shows the egocentric view used to localize the target Coke bottle, and the right frame shows the user executing the resulting reach instruction.

Navigator combines a live egocentric camera frame, persistent route context, and a structured prompt within a GPT Realtime pipeline to generate step-by-step guidance during movement. Figure 5 shows an example sequence in which the system guides the user toward a Coke bottle using the current camera view and the ongoing route state.

Its output follows a fixed guidance grammar of action, local anchor, embodied metric, check, and recovery, so each instruction is short, body-centered, and easy to execute in real time. The core mechanism is a guidance grammar that constrains the format of each output step:

$$\begin{aligned}\text{Step} = \{&\text{Action}\}\\ &+ \{\text{Local Anchor}\}\\ &+ \{\text{Embodied Metric}\}\\ &+ \{\text{Check}\}\\ &+ \{\text{Recovery}\}.\end{aligned}$$

Here, *Action* specifies the movement to perform, *Local Anchor* fixes the movement relative to the user's current body orientation, *Embodied Metric* expresses distance or direction in body-scaled units, *Check* specifies a verification action during motion, and *Recovery* specifies what to do if the expected condition is not met.

For example, Navigator can guide the user to the Coke bottle by saying: "Reach forward with your right hand, slightly to the right around 2 o'clock, within one arm's reach."

An example indoor walking step is shown below:

Action: walk forward

Local Anchor: keep the current body orientation

Embodied Metric: two steps
Check: sweep the cane after each step and stop if contact occurs
Recovery: if an obstacle is encountered, shift half a step to the right and try again

At runtime, the Navigator reads the next target from the room-route JSON, grounds it against the current live frame, and generates one guidance step in this grammar. If the live frame indicates that the scene still matches the planned route, the system advances to the next step. If the scene no longer matches, the Navigator revises the current step or requests a local rerouting update from the room-route representation. The resulting navigator output is then forwarded to the audio module.

4.7 Audio

The audio module converts the text output of Navigator and Toucher into spoken guidance. We implement text-to-speech with the OpenAI Audio API 'speech' endpoint using the 'gpt-4o-mini-tts' model in streaming mode (OpenAI, 2026a,c), so audio playback can begin before the full utterance is generated. In our implementation, the module receives candidate text from both branches, resolves conflicts, and then synthesizes a single spoken instruction for the user.

The system decides to speak when one of two conditions is met. First, Toucher output is emitted immediately when its priority score exceeds a predefined threshold, since these items correspond to nearby hazards or urgent checks. Second, Navigator output is emitted when the current route step changes, when a check or recovery clause becomes active, or when the user explicitly requests route guidance. If both branches produce output at the same time, Toucher takes precedence over Navigator, and Navigator is delayed until the urgent message has finished.

The length of each audio output is controlled before synthesis. Toucher messages are restricted to short utterances, typically one sentence, and only include the highest-priority item or top few tightly related items. Navigator messages are generated as a single guidance step under the guidance grammar, so each utterance contains only one action block at a time. This keeps the spoken output short enough to be executed before the next update cycle.

User speech input is captured through the microphone and transcribed with the OpenAI Audio API 'transcriptions' endpoint using the 'gpt-4o-mini-transcribe' model (OpenAI, 2026a,b). The transcribed text is treated as an interaction update and appended to the current Navigator prompt as additional user intent or clarification. If the user interrupts while audio is playing, playback is paused, the new transcription is processed, and the next Navigator step is regenerated using the updated prompt.

At runtime, the audio module is inserted after the Navigator and Toucher loops. Each update cycle collects branch outputs, applies the speech-decision policy, synthesizes audio if needed, and listens for new user speech in parallel. New user prompts are then fed back into the Navigator loop so that route generation remains conditioned on the latest user request.

4.8 Hardware

Figure 6 shows the physical prototype used to implement Touvigation. We built the system as a compact head-worn assembly so that the sensing direction remains aligned with the user's forward-facing body orientation, which is important for preserving the body-centric reference frame used throughout our guidance

Fig. 6. Hardware prototype of Touvigation shown from the front, side, top, and worn views. The prototype combines a head-mounted egocentric camera, a custom 3D-printed support structure, onboard Raspberry Pi computing, speech input, and audio output into a single wearable system.

model. The front, side, and top views show the structure of the device, while the worn view illustrates how the camera perspective follows the user's natural heading during movement.

The core sensing channel is an egocentric RGB camera mounted at the front of the headset. We use a head-mounted placement rather than a hand-held phone camera so that the visual stream stays stable relative to the user's body and does not depend on how the user happens to hold a device while walking, probing with a cane, or reaching for objects. This mounting choice is especially important for our body-centric interaction design because directional references such as clock-face descriptions are easier to interpret when the camera remains consistently aligned with the user's heading.

To support this form factor, we designed a custom 3D-printed housing and bracket system that fixes the camera and computation hardware into a rigid wearable frame. The 3D-printed structure serves three roles. First, it holds the camera at a repeatable angle so the system sees approximately what the user is facing. Second, it provides a stable mounting point for the onboard electronics without requiring the user to carry the sensing stack by hand. Third, it organizes the physical layout of the prototype so the full assembly can be worn as a single unit during navigation and object-search tasks. This mechanical design was important for moving from a desktop proof of concept to a deployable embodied prototype.

The onboard computation is handled by a Raspberry Pi, which functions as the local control hub for the wearable device. In our implementation, the Raspberry Pi interfaces with the camera, manages the live data stream, forwards multimodal input to the software pipeline, and coordinates speech input and audio playback. Using a Raspberry Pi allowed us to package the prototype as a self-contained mobile system rather

than relying on a tethered workstation, while still keeping the hardware stack inexpensive, reproducible, and easy to modify during iterative prototyping.

Finally, the workflow depends on three main I/O channels integrated into this prototype: the egocentric camera for continuous visual sensing, a microphone for capturing user speech and spoken queries, and an audio output channel for delivering navigation and toucher responses. Together, these components allow Touvigation to operate as a wearable embodied assistant that continuously perceives the scene, accepts natural-language user input, and returns timely spoken guidance during movement and interaction.

4.9 Demo

To illustrate Touvigation in use, we provide three supplementary videos that show goal-directed navigation in different indoor contexts. The first two videos are recorded in kitchen environments with different layouts and object configurations, while the third video is recorded in a hotel environment. Together, these examples show how the system supports navigation toward a user-specified goal while also providing hand-aware, near-range guidance during movement and interaction.

All spoken navigation instructions in the videos are in Chinese. We made this choice because our planned deployment and participant recruitment focus on blind users in China. To make the demonstrations understandable to a broader research audience, all three videos include English subtitles.

- **Kitchen scenario 1:** Supplementary video link
- **Kitchen scenario 2:** Supplementary video link
- **Hotel scenario:** Supplementary video link

5 Evaluation and Iteration Plan

At the time of writing, we completed the functional prototype and scenario-based demonstration videos, but we were not able to complete a full deployment evaluation with blind and low-vision participants within the project timeline. Because Touvigation is intended for real-time mobility and object interaction, a final evaluation with BLV users is a necessary next step before making claims about effectiveness, safety, or everyday usability. This section therefore describes the planned evaluation protocol, the rationale for the selected methods, and how the findings will guide subsequent design iterations.

5.1 Evaluation Goals

The evaluation will examine whether Touvigation's embodied guidance helps BLV users complete everyday indoor tasks more effectively than description-oriented assistance. Specifically, we will evaluate four questions:

- **Actionability:** Can users convert Touvigation's spoken guidance into immediate bodily action without extensive clarification?
- **Spatial grounding:** Do fixed body-centric anchors and embodied metrics reduce confusion compared with generic visual descriptions?
- **Safety and workload:** Does reachability-aware guidance reduce near-field hazards, unnecessary exploration, and perceived cognitive load?

- **Acceptability:** Do users find the wearable form factor, audio timing, and interaction style comfortable and appropriate for real-world use?

5.2 Participants and Recruitment

We plan to recruit 8–12 blind or low-vision adults for the evaluation. Participants will include people with diverse levels of vision impairment, mobility experience, and assistive-tool use, including cane users and users who have prior experience with AI-based visual assistance tools. We are working with collaborators at Tongji University in Shanghai to support local recruitment, study logistics, accessible testing arrangements, and participant safety planning. Recruitment will be conducted through blind-user communities, accessibility organizations, social media groups, and direct outreach to prior contacts who consent to being contacted again. All participants will provide informed consent and will be compensated for their time.

Because the current prototype uses spoken Chinese in the demonstration videos and is designed around the researcher's planned deployment context, the first evaluation will focus on Chinese-speaking BLV participants. This choice allows participants to evaluate the actual language, phrasing, and timing of the system rather than a translated approximation.

5.3 Study Design and Tasks

The study will use a mixed-methods within-subjects design. Each participant will complete a set of indoor tasks under two conditions: (1) Touvigation and (2) a baseline condition using a current description-oriented AI assistance workflow, such as a phone-based visual assistant. The order of conditions and tasks will be counterbalanced to reduce learning effects.

Tasks will be based on realistic scenarios used throughout the design process:

- **Room navigation:** move from a starting position to a target area in an unfamiliar indoor room.
- **Object finding:** locate a target object such as a trash can, chair, cup, or countertop item.
- **Near-field interaction:** approach and interact with an appliance or object surface, such as finding an air-conditioner control or confirming the edge of a table.
- **Recovery scenario:** respond to an unexpected obstacle or route mismatch and evaluate whether the system provides useful correction.

These tasks are chosen because they test the full interaction chain: route-level guidance from Navigator, near-field guidance from Toucher, body-centric reference frames, embodied distance units, and recovery instructions. The tasks also represent everyday contexts where scene descriptions alone are often insufficient.

5.4 Measures and Data Collection

We will collect both behavioral and experiential data. Behavioral measures will include task completion, completion time, number of navigation errors, number of system clarifications requested, number of researcher safety interventions, and observed near-miss events. System logs will capture generated instructions, timing, user speech input, and branch-level outputs from Navigator and Toucher.

Experiential measures will include post-task ratings of perceived actionability, spatial clarity, confidence, comfort, and cognitive workload. Participants will also complete a short post-study interview about confusing instructions, helpful guidance, trust, privacy concerns, hardware comfort, and situations where they

would or would not use the system. When appropriate, we will use standardized usability and workload instruments, such as SUS and NASA-TLX, alongside study-specific questions about embodied spatial guidance (Brooke, 1996, Hart and Staveland, 1988).

5.5 Safety Protocol

Because the evaluation involves mobility tasks with blind participants, safety will be prioritized over task performance. The study space will be pre-inspected, and a researcher will shadow participants at a close distance without providing task guidance unless intervention is needed. Participants will be allowed to use their normal mobility tools, including canes. The study will include a brief orientation, a practice trial, and clear stop rules. Any condition that creates discomfort, repeated confusion, or unsafe movement will be paused or terminated.

5.6 Analysis and Iteration

Quantitative results will be analyzed descriptively and, where sample size permits, with paired comparisons between Touvigation and the baseline condition. The main comparison will focus on whether Touvigation improves task completion, reduces clarification requests, and lowers perceived workload. Because BLV navigation behavior varies substantially across individuals, we will interpret aggregate results alongside per-participant patterns rather than relying only on statistical significance.

Qualitative data from observation notes and interviews will be analyzed thematically. We will focus on moments where participants misunderstood an anchor, rejected a distance unit, asked for repeated guidance, encountered a mismatch between instruction and reachable space, or adapted the system output into their own mobility strategy. These breakdowns are especially important because they reveal which parts of the guidance grammar and sensing envelope need revision.

The evaluation will directly inform the next design iteration. If users find clock-face references confusing in some postures, we will revise the reference grammar and add confirmation prompts. If step counts or cane-sweep estimates are inaccurate, we will adjust calibration and personalize embodied units. If Toucher interrupts too often, we will tune the priority scheduler and audio policy. If the headset is uncomfortable or socially inappropriate, we will revise the enclosure, mounting position, and audio output design. In this way, the evaluation is not only a validation step but also an iterative design mechanism for improving the prototype before longer-term field deployment.

6 Conclusions and Reflections

This project began from a clear problem: current AI-based visual assistance systems are often strong at generating descriptions, but much weaker at supporting immediate action for blind and low-vision users. The intended users in this work are not passive recipients of information. They navigate through practiced bodily strategies, tool use, and ongoing environmental interpretation. Based on that understanding, the central goal of this project was to move from description-oriented assistance toward embodied, actionable guidance. Touvigation addresses this goal by grounding system output in a fixed body-centric reference anchor, embodied movement units, and a reachability-aware sensing envelope. Rather than treating guidance as a generic scene summary, the project reframes navigation support as a process of helping users act in the moment with greater clarity, consistency, and confidence.

Several important insights emerged through the process. First, better perception alone does not automatically produce better assistance. What mattered most in our formative findings was not only whether the system could identify objects or layouts, but whether its output aligned with how users actually move, probe, and verify the world. Second, reference frames, units, and granularity are not minor interface details; they fundamentally shape whether guidance is usable. A technically correct instruction can still fail if it is not legible through the body. Third, the design process reinforced the importance of building from users' existing skills rather than expecting users to adapt to system-centric abstractions. This led to a stronger emphasis on embodied metrics, local verification, and progressive guidance rather than one-shot scene description.

More broadly, this project contributes to the field by offering both a design perspective and a system framework for future assistive AI. At the design level, it argues for a shift from helping blind and low-vision users know more about a scene to helping them do more within it. At the system level, it demonstrates how multimodal AI, wearable sensing, route planning, and near-range guidance can be combined around users' embodied interaction needs instead of around vision-first representations alone. In this sense, the project contributes not only a prototype, but also a transferable way of thinking about AI accessibility: effective assistive intelligence should be grounded in action, situated context, and the lived practices of the people it is meant to support.

7 Acknowledgement

We thank our collaborators at Tongji University in Shanghai for supporting the planned evaluation, including participant recruitment, local study preparation, and feedback on the prototype deployment context. We used AI-based tools to assist with parts of the system's code writing and implementation workflow. We also used AI tools during manuscript preparation for language polishing and grammar checking. However, AI was not used to generate the substantive content, claims, analysis, or core contributions of this paper.

References

Ahmetovic, D., Gleason, C., Ruan, C., Kitani, K., Takagi, H., and Asakawa, C. (2016). Navcog: a navigational cognitive assistant for the blind. In *Proceedings of the 18th international conference on human-computer interaction with mobile devices and services*, pages 90–99.

Ahmetovic, D., Oh, U., Mascetti, S., and Asakawa, C. (2018). Turn right: Analysis of rotation errors in turn-by-turn navigation for individuals with visual impairments. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 333–339.

Apple (2023). Apple previews live speech, personal voice, and more new accessibility features. Accessed: 2026-04-04.

Azaria, A. and Mitchell, T. (2023). The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976.

Azenkot, S., Wobbrock, J. O., Prasain, S., and Ladner, R. E. (2012). Input finger detection for nonvisual touch screen text entry in perkinput. In *Proceedings of graphics interface 2012*, pages 121–129.

Be My Eyes (2023). Introducing: Be my ai. Accessed: 2026-04-04.

Be My Eyes (2026). Accessibility technology for blind & low vision people. Accessed: 2026-04-04.

Bharadwaj, A., Shaw, S. B., and Goldreich, D. (2019). Comparing tactile to auditory guidance for blind individuals. *Frontiers in Human Neuroscience*, 13:443.

Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., Miller, R., Tatarowicz, A., White, B., White, S., et al. (2010). Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342.

Brooke, J. (1996). Sus: A quick and dirty usability scale. In Jordan, P. W., Thomas, B., McClelland, I. L., and Weerdmeester, B., editors, *Usability evaluation in industry*, pages 189–194. Taylor & Francis.

Chaudary, B., Pohjolainen, S., Aziz, S., Arhippainen, L., and Pulli, P. (2023). Teleguidance-based remote navigation assistance for visually impaired and blind people—usability and user experience. *Virtual Reality*, 27(1):141–158.

Coughlan, J. M. and Shen, H. (2013). Crosswatch: a system for providing guidance to visually impaired travelers at traffic intersection. *Journal of assistive technologies*, 7(2):131–142.

Dourish, P. (2001). *Where the action is: the foundations of embodied interaction*. MIT press.

Fiannaca, A., Apostolopoulous, I., and Folmer, E. (2014). Headlock: a wearable navigation aid that helps blind cane users traverse large open spaces. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*, pages 19–26.

Froese, T. and Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial intelligence*, 173(3-4):466–500.

Google (2026). Ml kit pose detection. Accessed: 2026-04-05.

Google LLC (2025). Lookout - assisted vision. Google Play. Accessed: 2026-04-04.

Guerreiro, J., Sato, D., Asakawa, S., Dong, H., Kitani, K. M., and Asakawa, C. (2019). Cabot: Designing and evaluating an autonomous navigation robot for blind people. In *Proceedings of the 21st international ACM SIGACCESS conference on computers and accessibility*, pages 68–82.

Hart, S. G. and Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in Psychology*, volume 52, pages 139–183. Elsevier.

Holloway, L., Ananthanarayan, S., Butler, M., De Silva, M. T., Ellis, K., Goncu, C., Stephens, K., and Marriott, K. (2022a). Animations at your fingertips: using a refreshable tactile display to convey motion graphics for people who are blind or have low vision. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–16.

Holloway, L. M., Goncu, C., Ilsar, A., Butler, M., and Marriott, K. (2022b). Infosonics: Accessible infographics for people who are blind using sonification and voice. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Hutchins, E. (1995). *Cognition in the wild*. MIT Press.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026.

Klemmer, S. R., Hartmann, B., and Takayama, L. (2006). How bodies matter: five themes for interaction design. In *Proceedings of the 6th conference on Designing Interactive systems*, pages 140–149.

Liu, R., Zhang, J., Schön, A., Müller, K., Zheng, J., Yang, K., Guo, A., Gerling, K., and Stiefelhagen, R. (2024). Objectfinder: An open-vocabulary assistive system for interactive object search by blind people. *arXiv preprint arXiv:2412.03118*.

Liu, T., Hernandez, J., Gonzalez-Franco, M., Maselli, A., Kneisel, M., Glass, A., Chudge, J., and Miller, A. (2022). Characterizing and predicting engagement of blind and low-vision people with an audio-based navigation app. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7.

Malaekah, E., Alfahad, O., Bakouri, M., Gadallah, A., Selman, S., Al Rashdi, A., and Saied, H. (2026). Sound-based navigation system for visually impaired individuals. *Journal of Radiation Research and Applied Sciences*, 19(1):102160.

Messaoudi, M. D., Menelas, B.-A. J., and Mcheick, H. (2022). Review of navigation assistive tools and technologies for the visually impaired. *Sensors*, 22(20):7888.

Microsoft Accessibility Blog (2017). Seeing ai app is now available in the ios app store. Accessed: 2026-04-04.

Microsoft Research (2026). Microsoft soundscape. Accessed: 2026-04-04.

OpenAI (2026a). Realtime api voice design. Accessed: 2026-04-05.

OpenAI (2026b). Speech-to-text. Accessed: 2026-04-05.

OpenAI (2026c). Text-to-speech. Accessed: 2026-04-05.

OpenCV (2026). Stitching detailed panorama. Accessed: 2026-04-05.

OrCam (2024). Orcam myeye 3 pro. Accessed: 2026-04-04.

Penuela, R. E. G., Jung, C., Lin, S. Y., Hu, R., and Azenkot, S. (2026). How multimodal large language models support access to visual information: A diary study with blind and low vision people. *arXiv preprint arXiv:2602.13469*.

Sato, D., Oh, U., Guerreiro, J., Ahmetovic, D., Naito, K., Takagi, H., Kitani, K. M., and Asakawa, C. (2019). Navcog3 in the wild: Large-scale blind indoor navigation assistant with semantic features. *ACM Transactions on Accessible Computing (TACCESS)*, 12(3):1–30.

Seo, J., Xia, Y., Lee, B., Mccurry, S., and Yam, Y. J. (2024). Maidr: Making statistical visualizations accessible with multimodal data representation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–22.

Siu, A. F., Sinclair, M., Kovacs, R., Ofek, E., Holz, C., and Cutrell, E. (2020). Virtual reality without vision: A haptic and auditory white cane to navigate complex virtual worlds. In *Proceedings of the 2020 CHI conference on human factors*

in computing systems, pages 1–13.

Southern, C., Clawson, J., Frey, B., Abowd, G., and Romero, M. (2012). An evaluation of brailletouch: mobile touchscreen text entry for the visually impaired. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*, pages 317–326.

Suchman, L. A. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge University Press.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9(4):625–636.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., and Zhao, H. (2024). Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911.

Zeraati, F. Z., Cao, Y. T., Qiao, Y., Daumé III, H., and Kacorri, H. (2026). Say it my way: Exploring control in conversational visual question answering with blind users. *arXiv preprint arXiv:2602.16930*.

Zhang, T. and Lillianfeld, L. (2024). Talkback uses gemini nano to increase image accessibility for users with low vision. Android Developers Blog. Accessed: 2026-04-04.

Zhao, Y., Bennett, C. L., Benko, H., Cutrell, E., Holz, C., Morris, M. R., and Sinclair, M. (2018). Enabling people with visual impairments to navigate virtual reality with a haptic and auditory cane simulation. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–14.