# Designing for Ring Camera Use: Spatial Interaction Strategies of Blind and Low Vision People


Ayaka Tsutsui
University of Tsukuba
Tsukuba, Japan
Miraikan – The National Museum of Emerging Science and Innovation
Tokyo, Japan
ayakatsutsui1128@digitalnature.slis.tsukuba.ac.jp

Xiyue Wang
Miraikan – The National Museum of Emerging Science and Innovation
Tokyo, Japan
wang.xiyue@lab.miraikan.jst.go.jp

João Guerreiro
LASIGE, Faculdade de Ciências
Universidade de Lisboa
Lisbon, Portugal
jpguerreiro@fc.ul.pt

Hironobu Takagi
Miraikan – The National Museum of Emerging Science and Innovation
Tokyo, Japan
hironobu.takagi@jst.go.jp

Yoichi Ochiai
University of Tsukuba
Tsukuda, Japan
wizard@slis.tsukuba.ac.jp

Chieko Asakawa
Graduate School of Science and Technology, Keio University
Kanagawa, Japan
casakawa@keio.jp



**Figure 1: Participants exploring exhibits using the ring camera with body movements. (a) For wall-mounted exhibits, the participant keeps distance from the exhibit while using the right hand as an anchor. (b) For horizontally placed exhibits, the participant uses the right hand as an anchor to maintain balance. (c) For horizontally placed wide exhibits, the participant engages the whole body in exploration.**


## Abstract

Blind and low vision (BLV) users rely on touch to build spatial understanding of the physical world. However, much information cannot be accessed through touch alone. Camera-based assistive systems help bridge this gap, yet how BLV users coordinate their hands and bodies across tactile and non-tactile media remains unclear. We investigate the spatial interaction strategies that BLV users employ when accessing information through a finger-worn ring camera. Eleven BLV participants accessed four media that varied in tactile cues through a Wizard-of-Oz system, which delivered information audibly at multiple granularities. Participants consistently used one hand as an anchor while adjusting the camera-wearing hand three-dimensionally, with the anchor serving as both a spatial reference and a postural support. Retrieval spanned scales from torso-level framing to fingertip-level calibration, yet intermediate-range capture remained particularly challenging. We discuss design implications for ring camera-based systems that support BLV users' spatial interaction with diverse media.



## CCS Concepts

• **Human-centered computing** → **Empirical studies in accessibility**; *Accessibility systems and tools.*

## Keywords

Wearable devices, Ring-shaped camera, Blind and low vision users, Tactile exploration, Information seeking, Body movement

**ACM Reference Format:**
Ayaka Tsutsui, Xiyue Wang, João Guerreiro, Hironobu Takagi, Yoichi Ochiai, and Chieko Asakawa. 2026. Designing for Ring Camera Use: Spatial Interaction Strategies of Blind and Low Vision People. In *The 28th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '26),*




*ASSETS '26, Vila Nova de Gaia, Portugal*



ACM ISBN 979-8-4007-2521-0/2026/10
https://doi.org/10.1145/3797867.3828989

*October 25–28, 2026, Vila Nova de Gaia, Portugal.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3797867.3828989

# 1 Introduction

For blind and low vision (BLV) users, touch plays a central role in perceiving shape and layout, supporting content understanding and spatial awareness [34, 37]. Reflecting this, Human-Computer Interaction (HCI) research has developed a wide range of tactile representations, including 2D tactile graphics [22, 24], 3D models [23, 39], and interactive 3D models augmented with audio feedback [12, 46], all of which are designed so that BLV users can access information through touch. Beyond the representations themselves, prior work has also examined how BLV users engage with them. For example, on large tactile surfaces, one hand is placed as an anchor while the other explores [20], in 2D tactile graphics, both hands are used in coordinated simultaneous tracing [63], and broader classification is followed by finer identification in a staged manner [50].

However, much of the information BLV users need lies beyond touch, including wall-mounted signs and printed documents [10]. Camera-based assistive systems have therefore been widely used to access such non-tactile information. Smartphone applications such as SeeingAI [35] and WorldScribe [9] provide scene descriptions based on the camera view, enabling BLV users to access visual information. These systems, however, require users to pause their hand movements to capture an image, making them difficult to use concurrently with exploration [16]. Head-worn smart glasses have been explored as a hands-free alternative [29, 31]. Despite this advantage, their eye-level viewpoint is poorly suited to examining objects being explored by hand and may result in unnatural postures as users orient their heads toward the target [56].

To bring the camera's viewpoint closer to the user's hand, finger-worn ring cameras have been proposed, which combine a camera with audio feedback to support information access through pointing at or tracing targets with the finger [17, 41, 52, 55]. These studies have positioned the ring camera as a device for reading text at the fingertip. In our previous work, we observed BLV users exploring tactile-rich 3D media with a ring camera [58]. Compared with smartphones, the ring camera supported smoother two-handed exploration, with one hand often serving as an anchor while the other explored. However, it remains unclear whether these anchoring and hand-body coordination strategies extend to interactions without reliable tactile reference points.

This gap has important implications for the design of ring camera-based systems. Research on bodily perception has shown that the choice of where to touch and how to move is itself part of how information is acquired [18, 27]. Building on this work, we use the same ring camera hardware [58] as a research probe for examining embodied information-access strategies across public information media, which we define as physical materials presenting printed text, tactile features, or both.

Based on this perspective, we investigate what bodily strategies BLV users employ when accessing information with a ring camera to inform the design of future ring camera-based systems and their interactions. Specifically, we address two research questions (RQs):

- **RQ1**: What hand movements and body postures emerge when BLV users access information through a ring camera across media that differ in the availability of tactile cues?
- **RQ2**: How do these movements and postures correspond to the types of information obtained and the outcomes of information access?

To investigate these strategies, we used museum exhibits as a concrete setting where information is distributed across diverse media with varying tactile affordances, requiring active bodily exploration [2, 33]. This allowed us to observe how BLV users adapted their hand and body movements across media types within a single environment. We conducted a user study where eleven BLV participants accessed physical media using a ring camera paired with a Wizard-of-Oz audio system (Figure 1 (a)–(c)). The study included four media conditions spanning a gradient of tactile availability: (1) a tactile-rich display combining a 2.5D object with an adjacent text panel, (2) a wall-mounted panel with text only, (3) a wall-mounted panel with subtle tactile cues, and (4) a printed catalog placed on a horizontal surface (Figure 3 (a)–(d)).

Our analysis revealed a consistent bodily pattern in BLV users' information-access behavior (RQ1). Users placed one hand on the medium as an anchor while adjusting the camera hand three dimensionally, and this structure emerged across all media regardless of the availability of tactile cues. The anchoring hand served two distinct roles, as a spatial reference that indicated where each piece of text was located, and as a postural support that stabilized camera orientation and body balance. These bodily strategies corresponded to different granularities of information retrieval, with broad framing that mobilized the arm and torso for capturing overall layout, and fine-grained adjustments at the fingertip for reading detailed text (RQ2). The success of information access, however, depended on whether the anchoring hand could fulfill its spatial-reference role: when the medium lacked tactile cues, the anchoring hand functioned only as postural support, and participants had to rely on memory during exploration, leading to increased cognitive load and notably more difficult access.

Our contributions are as follows:

- We systematically characterize the repertoire of hand and body movements that BLV users employ when accessing information through a ring camera across physical media that differ in the availability of tactile cues.
- We show that the anchoring hand functions as both a spatial reference and postural support, acting as an organizing axis for information access rather than a passive anchor.
- Drawing on these observations, we derive design implications for ring camera-based systems that support BLV users' bodily strategies regardless of the presence of tactile cues.

# 2 Related Work

## 2.1 Embodied Exploration by BLV People

Information access for BLV users can be understood not as a passive act of receiving information by ear, but as an active exploration mediated through the motion of the hands and fingers. This view is grounded in Gibson's characterization of touch as an active sense [18], and in Lederman et al.'s finding that people use stereotypical hand movements tuned to the property they seek

to perceive [27]. Together, these findings make clear that haptic perception is inseparable not only from what is touched but also from how the hand touches and moves. This view extends to assistive technologies for BLV users. El-Glaly et al. showed that BLV users can read text more effectively by touching a touchscreen [14]. Similarly, ImageAssist [40] and ImageExplorer [28] demonstrated that touch-based image exploration conveys spatial structure that textual descriptions alone cannot. Together, these studies highlight the value of accessing visual information through bodily motion.

Building on this view, studies have examined how BLV users coordinate their fingers and hands during exploration. On large touchscreens, Guerreiro et al. found that participants used one hand as an anchor for the other [20]. In 2D tactile graphics, Zhao et al. found that simultaneous two-handed tracing led to deeper understanding than single-handed tracing [63], while Sharma et al. observed mirrored movements for symmetrical shapes and a common progression from broad classification to finer identification [50]. For 3D models, guided two-handed exploration has similarly been shown to support deeper understanding [57].

These studies consistently show that exploration involves local tactile perception, two-handed coordination, anchor-hand use, and structured shifts from coarse to fine inspection. However, much of the information BLV users encounter is beyond direct touch, including printed text on walls, captions accompanying 3D objects, and printed documents [10]. It remains unclear whether and how these strategies extend from the fingers and hands to the arms and whole body in such contexts. Our study addresses this gap by examining how bodily strategies emerge when BLV users access information through a ring camera.

## 2.2 Wearable and Camera-Based Systems for BLV Users

Visual-assistance systems based on smartphone have been widely developed as assistive technologies through which BLV users can access physical media. VizWiz [5] is a system in which remote sighted workers answer questions about images captured by BLV users. Commercial applications such as SeeingAI [35] have since integrated multiple features, including OCR, scene description, and object recognition. More recently, WorldScribe [9] has provided live visual descriptions that adapt to camera motion and the captured visual content. While smartphones are widely used in the daily life of BLV users [49], it has been repeatedly pointed out that aiming the camera at a target, and adjusting distance are difficult [10, 26].

As an alternative to handheld smartphones, head-worn cameras have been widely explored. Commercial smart glasses such as Envision Glasses [15] and OrCam [43] attach a camera to the frame and provide OCR and scene description, enabling hands-free access to information. In the research literature, a pedestrian detection study showed that a head-worn camera allows users to naturally adjust the field of view by moving the head [31]. A subsequent remote usability study using similar smart glasses for BLV users reported that the captured field of view varies substantially with bodily movement and the direction of sound sources [29]. In addition, Teng et al. reported that, because the viewpoint of a head-worn camera is fixed to the eye line, BLV users who typically recognize targets through the hand experience an unnatural posture when orienting the head toward the target [56].

In response to these issues, finger-mounted cameras have brought the camera viewpoint closer to the user's hand [53]. As an early example, EyeRing [41] paired a ring-shaped camera with audio feedback, allowing BLV users to point at a target to retrieve information. Building on this approach, FingerReader [6, 52] enabled BLV users to read printed text sequentially by tracing text lines with a finger. HandSight [17, 55] combined a finger-worn camera with haptic and auditory directional guidance, refining the coupling between finger movement and text reading. These studies focus on fingertip movement in text reading and target pointing.

More recently, work has begun to extend the scope of finger-worn cameras beyond text to diverse types of media. In our prior work, we investigated tactile exploration support using a finger-worn camera across multiple three-dimensional objects [58]. However, that study primarily focused on directly touchable targets, leaving unclear how finger-worn cameras are used in contexts where touchable and non-touchable media coexist. In addition, much of the existing work has focused on technical aspects such as the placement of the camera and its recognition accuracy [53], and how BLV users deploy bodily strategies that mobilize not only the fingertip but also the arm and the whole body has not been sufficiently investigated.

Our study builds on the lineage of existing wearable camera systems, but shifts the focus from technical implementation to the observation of BLV users' bodily strategies. Through this perspective, our aim is to clarify how BLV users integrate a wearable camera with their body during the exploration of diverse physical media, ranging from text to three-dimensional objects.

# 3 User Study

We conducted a Wizard-of-Oz (WoZ) user study with eleven BLV participants who used a ring camera to access four museum media differing in tactile affordances. This setting enabled us to observe how hand and body strategies adapted to different forms of information access.

## 3.1 Ring Camera Setup

We used the same ring camera prototype as in our prior work [58]. The ring camera is worn on the proximal segment of the left index finger. Figure 2 shows the device and how it is worn (a)–(c) and dimensional measurement renderings generated from its CAD model (d)–(f). This placement on the non-dominant hand follows prior findings that BLV users prefer wearing assistive devices on the non-dominant hand [58]. This configuration enables users to move their hands independently while accessing information across media, including those with and without tactile features.

*3.1.1 Device Design.* The ring camera consists of a finger-worn ring module and a wrist-mounted processing module, connected by wires. The ring module (Figure 2 (b) and (e)) weighs 6.4 g and measures 29.3 mm × 28.4 mm, with a thickness of 8.4 mm. Two variants (16 mm and 22 mm inner diameters) accommodate different finger sizes. Its outer shell is made of anodized aluminum with a polycarbonate inner lining for wearing comfort.

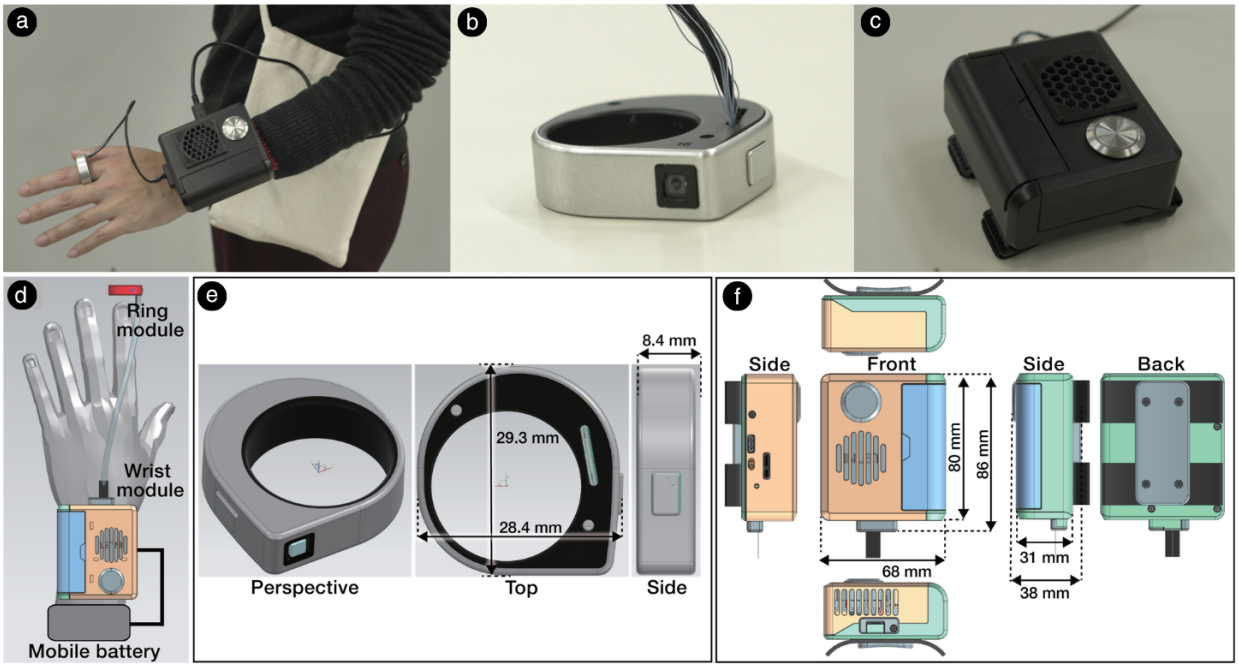


**Figure 2: The same ring camera and CAD model from [58]. (a) The ring is worn on the proximal segment of the left index finger, with the wrist module attached to the left wrist. The connected power bank is stored in a small neck bag. (b) Close-up of the physical ring module. (c) Close-up of the physical wrist module. (d) CAD rendering of the ring-camera device in its worn configuration. (e) Close-up CAD rendering of the ring module. (f) Close-up CAD rendering of the wrist module.**

The wrist module (Figure 2 (c) and (f)) weighs 135 g and is secured with two Velcro straps. It measures 86 mm (L) × 68 mm (W) × 38 mm (H). The current prototype is powered by an external 5,000 mAh power bank (via USB-C), providing approximately 3 hours of operation, which users carry in a small pouch to avoid restricting movement (Figure 2 (a)) . The ring module captures images and transmits them to the wrist module, where they are processed and streamed to a laptop computer (MacOS) via Wi-Fi. The details of the two modules are described below.

(1) *Ring module.* A miniature camera captures 0.7 MP frames at 920 × 736 resolution (10 fps) in RAW 10-bit format, with a focal length of 0.574 mm and fields of view of 104.7° (horizontal), 91.5° (vertical), and 120.0° (diagonal).
(2) *Wrist module.* The wrist module includes a Zynq UltraScale+ MPSoC (XCZU5EG), wireless communication, a battery slot, and a cooling fan. An onboard ISP converts RAW frames into 920 × 736 YUV422 8-bit images, which are transmitted to a laptop via Wi-Fi (ATWLC3000-MR110CA).

## 3.2 Participants

Eleven BLV participants were included in the study, as summarized in Table 1 (six male, five female; mean age = 43.9, SD = 16.6). Participants were recruited through a national newsletter for people with visual impairments. Five participants (P3 and P8–P11) had low vision, and the remaining six (P1, P2, and P4–P7) were blind. All participants were right-handed and regularly used a white cane.

All participants reported regularly accessing information from both textual and tactile media in their daily lives. They read printed materials such as pamphlets, manuals, and official documents, typically with the support of smartphone-based assistive applications (Table 1). All participants commonly used SeeingAI [35], and also reported using other applications such as SwiftAI [61] and Be My AI [3], with usage experience ranging from less than one year to more than five years. Participants also interacted with tactile media such as maps, 3D models, and museum exhibits, with an average of 14.1 years of tactile exploration experience (SD = 14). Those with later-onset visual impairment reported shorter histories of tactile exploration and greater reliance on auditory information.

## 3.3 Experimental Setup

Participants interacted with the exhibits using the ring camera introduced earlier. Two ring sizes were available to accommodate different finger sizes. When needed, hypoallergenic medical tape was used to improve stability. Audio descriptions were pre-recorded and played through a neck-worn Bluetooth speaker (Sharp AN-SS3).

## 3.4 Exhibits and Materials

We used four science-related materials: three museum exhibits and one museum catalog. Following prior work [57], we selected materials spanning different tactile affordances, from text-dominant materials with few or no tactile cues to tactile-rich exhibits combining physical features and textual descriptions. To examine the influence of bodily movement, we also included different spatial

| ID | Age | Gender | Visual status | Age of onset | Years of tactile exploration | Braille literacy | Type of application used |
|---|---|---|---|---|---|---|---|
| P1 | 56 | F | Blind | 0 | 50 | Yes | SeeingAI, Swift AI |
| P2 | 61 | F | Blind | 51 | 2 | No | SeeingAI, Be My AI |
| P3 | 43 | M | Low vision (visual acuity: 0.2) | 26 | 5 | No | SeeingAI, Swift AI |
| P4 | 22 | M | Blind | 3 | 19 | Yes | SeeingAI |
| P5 | 50 | M | Blind | 20 | 10 | No | SeeingAI, Be My AI |
| P6 | 23 | F | Blind | 1 | 17 | Yes | SeeingAI, Be My AI |
| P7 | 60 | M | Blind | 55 | 5 | No | Be My Eyes |
| P8 | 53 | M | Low vision (central scotoma) | 43 | 8 | No | SeeingAI |
| P9 | 33 | F | Low vision (visual acuity: right 0.04, left 0.02) | 0 | 23 | Yes | SeeingAI, shikAI, Be My AI |
| P10 | 20 | M | Low vision (visual acuity: right 0.02, left 0.04) | 0 | 15 | Yes | Magnification, Color inversion, SeeingAI |
| P11 | 62 | F | Low vision (visual acuity: 0.2) | 58 | 1 | No | SeeingAI |

**Table 1: Details of the eleven BLV participants in the user study.**

configurations, including wall-mounted and horizontal displays. All materials contained text in a primary language, and some included translations in two additional languages.

**Exhibit A: Fetal Development Exhibit** shown in Figure 3 (a), presents the developmental stages of a fetus and consists of two sections. The first is a 54 cm × 54 cm wall-mounted display containing a 2.5D fetus model. Beneath the model, a 36 cm (W) × 6 cm (H) area presents non-embossed text descriptions in 16-point font. The second is a 24 cm (W) × 19 cm (H) horizontally arranged display with text descriptions at the top. Below the text, eight circular elements, each 9 cm in diameter, are arranged left to right, each containing the number of days in the developmental stage and a 2.5D model showing the fetus's actual size at that stage.

**Exhibit B: Nobel Prize Laureate Panel** shown in Figure 3 (b), is a wall-mounted informational panel describing the research of Nobel Prize laureate Masatoshi Koshiba[1]. The panel measures 65 cm (W) × 28 cm (H) and presents all information as non-embossed text in approximately 16-point font.

**Exhibit C: Cell Structure Panel** shown in Figure 3 (c), is a wall-mounted panel explaining the structure of a cell. The panel measures 80 cm (W) × 50 cm (H) and presents text in 80-point font, with slight embossing on both the text and the surrounding frame.

**Exhibit D: Catalog Page** shown in Figure 3 (d), is a page from a science museum catalog describing five key initiatives for shaping the future through science and technology. The page, A4 in size (21 cm × 29.7 cm), is placed on a horizontal surface and presents all information as non-embossed text in 10-point font.

## 3.5 Wizard-of-Oz System

We employed the Wizard of Oz (WoZ) methodology to elicit design requirements for the proposed system [11]. The WoZ setup allowed us to examine participants' exploration and interaction strategies independently of technical factors such as recognition accuracy, processing latency, and detection failures, rather than evaluating the performance of an automated visual recognition system.

The WoZ methodology is widely adopted in interaction design research to experimentally evaluate user experiences with technically incomplete systems [1, 44]. In our study, a researcher acted as the "Wizard," providing audio descriptions and system feedback based on the content captured by participants' cameras. Examples of ring camera views appear in Figure 7 in Appendix. The Wizard selected and played appropriate audio descriptions from a pre-prepared set using a dedicated interface designed for each exhibit.

The Wizard triggered audio descriptions only after participants reached a stable hover state over the medium, not while the camera was moving. Internal testing kept response times below 500 ms, and our blind co-author reported no disruptive delay. This procedure followed established WoZ guidelines for reducing human-in-the-loop confounds [13]. Participants were unaware of the Wizard's involvement and perceived all feedback as automatically generated.

*3.5.1 System Output.* The system provided participants with three types of auditory feedback.

**Capture Success Chime.** A short chime indicated successful capture, followed immediately by the corresponding audio description, allowing participants to relax their camera-holding posture.

**Distance Warning.** A voice message ("too far") indicated that the camera was too distant from the target.

**Camera Obstruction Alert.** A beep signaled that the camera was obstructed, prompting participants to clear the view.

*3.5.2 Unsuccessful Capture State.* When the camera did not capture sufficient information to trigger a predefined description, the system produced no output. Based on feedback from a blind co-author during WoZ protocol development, we avoided continuous corrective cues (e.g., error earcons), which could unnecessarily increase cognitive load [25]. The researcher intervened only when participants appeared unable to continue, for example by prompting them to adjust the camera or informing them that no predefined content was in view, without revealing task-specific information.

*3.5.3 Audio Description Structure.* Depending on the captured content, audio descriptions were provided at three levels of granularity.

**Overview descriptions** were provided when the camera captured the entire exhibit. As the broadest level in the hierarchy, they outlined the overall theme of the exhibit and the spatial arrangement of its sections. While the overview enabled participants

[1]https://www.nobelprize.org/prizes/physics/2002/koshiba/facts/

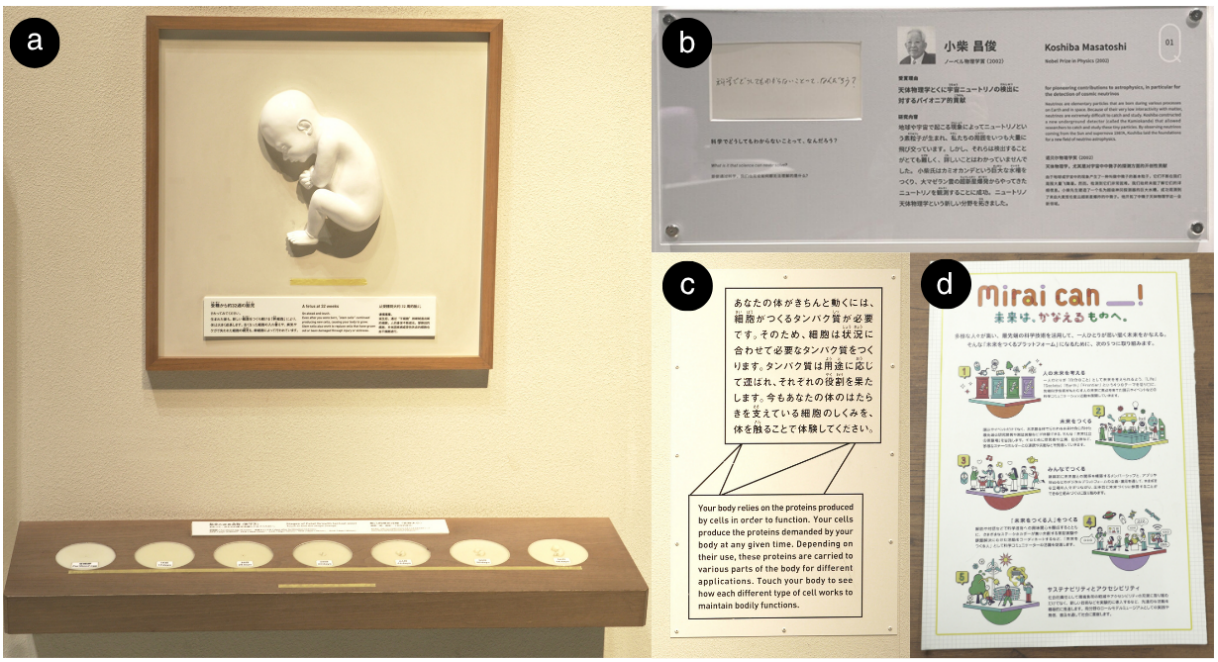


**Figure 3: Four exhibits used in the user study. (a) Exhibit A: a two-part display combining a wall-mounted section and a horizontal section, with tactile elements that can be felt by hand. (b) Exhibit B: a wall-mounted panel without tactile texture. (c) Exhibit C: a wall-mounted panel with limited tactile texture. (d) Exhibit D: a horizontally placed page without tactile texture.**

to grasp the overall structure, understanding the content of each specific section required bringing the camera closer.

**Section descriptions** provided an overview of a specific area within the exhibit, such as the title, subtitle, and explanatory text within that section. To understand the specific content, participants needed to bring the camera closer.

**Detail descriptions** were the lowest level of the hierarchy and provided concrete content, such as individual text and illustrations. Capturing an illustration triggered a description of its visual content, while text caused the corresponding passage to be read aloud.

The number of descriptions prepared for each exhibit varied by granularity level. Exhibit A included 3 Overview, 3 Section, and 18 Detail descriptions; Exhibit B included 1 Overview, 4 Section, and 15 Detail descriptions; Exhibit C included 1 Overview, 2 Section, and 2 Detail descriptions; and Exhibit D included 1 Overview, 5 Section, and 13 Detail descriptions (Total 68 target locations).

*3.5.4 Information Scope and Consistency.* The information content of the audio descriptions was determined in accordance with prior work [57]. Descriptions were limited to information explicitly presented in the exhibits, and no supplementary or contextual information beyond what was visually displayed was provided.

Although the framing and coverage of captured images inevitably varied across participants, consistency in the audio descriptions was ensured. All audio descriptions were pre-authored for discrete information units at the three hierarchical levels (Section 3.5.3), and the Wizard's judgment was restricted to a binary decision of whether a captured image contained a particular information unit. When the same information unit was identified, the audio description was played regardless of framing, ensuring uniformity across participants. For textual content, playback always started from the beginning of the passage, minimizing inconsistencies due to individual differences in capture position.

*3.5.5 Participant Input.* When participants wished to interrupt an ongoing audio description, they verbally uttered "stop," upon which the Wizard ceased playback.

## 3.6 Task and Procedure

The entire session took approximately 2.5 hours, including scheduled breaks. After completing the pre-study interview on demographic background and prior experiences, participants received training on how to operate the ring camera. During this training, participants practiced understanding the relationship between the position of the hand and the area captured by the camera. To avoid introducing bias, they were not informed about the types or levels of information that would be provided by the system.

In the main task, participants explored the four exhibits described in Section 3.4, with the presentation order counterbalanced across participants. Each exhibit was allotted up to 20 minutes, and participants were encouraged to use a think-aloud protocol throughout. Before each trial, participants first touched the boundaries and overall placement of the exhibit to establish a safe starting reference and gain an approximate sense of its spatial layout. The task for each exhibit consisted of two phases designed to capture different aspects of exploratory interaction.

In **Phase 1 (free exploration)**, participants explored the exhibit using both touch and the ring camera to understand its overall content. This phase was designed to observe spontaneous exploration strategies and hand movement patterns without imposing specific information goals. Each exploration lasted up to 10 minutes, and participants could stop earlier if they felt they had sufficiently understood the exhibit. After Phase 1, participants rated the perceived difficulty of the exploration using the Single Ease Question (SEQ, 1 = very difficult, 4 = neutral, 7 = very easy) [48].

In **Phase 2 (goal-directed exploration)**, participants were asked to answer three questions about each exhibit. Because participants had already explored the exhibit in Phase 1, this phase was not intended to measure fresh information seeking. Instead, it was designed to examine goal-directed retrieval based on knowledge acquired during free exploration. Participants were instructed to re-explore the exhibit to locate and confirm information needed to answer each question, even if they already knew the answer.

The three questions progressively elicited different levels of information acquisition, including grasping the **overall content**, locating **section-level content**, and identifying **detailed information** (Table 7, Appendix). The order of the questions was counterbalanced across participants. This phase allowed us to examine how participants adapted their hand movements when retrieving previously explored information at different levels of granularity. After answering each question, participants answered the SEQ.

After completing both phases for each exhibit, participants rated the usability of the ring camera on three aspects: the ease of grasping the overall content, locating relevant sections, and accessing detailed information. Each item was rated on a 7-point Likert scale (1 = very difficult, 4 = neutral, 7 = very easy), as shown in Table 2.

After completing the tasks, participants rated a set of 7-point Likert items (1 = strongly disagree, 4 = neutral, 7 = strongly agree) and completed a semi-structured interview. The Likert items (Table 3) assessed three aspects: (1) the spatial relationship between hand position and camera coverage; (2) the appropriateness of the system's audio feedback relative to the level of information sought; and (3) the extent to which participants consciously adjusted their hand movements for different exploration goals. The semi-structured interview further explored participants' exploration strategies, perceived differences across exhibits with varying tactile and spatial properties, and views on the advantages, and potential applications of the ring camera (Table 4). Procedures were approved by the institution's ethics committee, and participants provided informed consent, receiving a compensation of approximately 50 USD.

## 3.7 Data Collection and Analysis

All sessions were video recorded, including task performance (bodily actions and think-aloud) and pre and post experiment interviews. Images captured by the ring camera and the audio output provided by the Wizard were logged and used for analysis. To analyze participants' bodily actions, we divided task activity into two steps: 1) positioning the ring camera to retrieve information and 2) listening while audio feedback was played. We focused on these phases, as they most clearly reflect interaction with the device. When participants were unable to aim the camera or ceased movement, a researcher intervened without providing task-related information.

Based on the video analysis and the coding scheme established in prior work [57, 58], we identified the 15 bodily action codes shown in Table 5. Using these codes, we analyzed 748 observation instances collected from eleven participants. Of the 748 instances, 549 contained at least one bodily action code in either stage. Because multiple codes could be assigned to a single instance, these 549 instances yielded 734 code occurrences: 564 during camera positioning and 170 during audio playback. Table 6 reports the frequencies of the 15 codes across the four exhibits. Instances in which participants did not raise the ring camera or were not attending to the relevant medium were classified as "none" and excluded from the bodily action frequency counts.

Although our study consisted of two phases, we did not separate the analysis by phase; instead, we conducted the analysis in the following two steps. Coding was performed independently by two authors, after which the other authors reviewed the coding scheme and, through discussion, finalized it. We then organized the codes into higher-level themes using affinity diagramming [4]. Participants completed the SEQ after both the free-exploration and question phases. All interviews were transcribed, and we performed a hybrid deductive-inductive approach, following what Braun and Clarke [7] refer to as codebook thematic analysis. The resulting themes were integrated with the results of the bodily action coding.

# 4 Results

Across the study, participants used the ring camera to freely explore each medium, accessing most of the system-provided information while developing an understanding of its layout. They valued being able to independently choose what to access and retrieve complete information, contrasting this with prior experiences using audio guides, where content was often abbreviated. At the same time, they encountered challenges in regulating camera distance, adapting body posture, and constructing a spatial understanding of media lacking tactile cues. We detail these findings below.

## 4.1 Overview of Observed Behaviors

Across all tasks, the mean completion times were 5.9 minutes (SD = 2.6) for Exhibit A, 5.9 minutes (SD = 1.6) for Exhibit B, 2.4 minutes (SD = 1.2) for Exhibit C, and 7.2 minutes (SD = 1.6) for Exhibit D. Overall, participants consistently used one hand as an anchor, while employing additional bodily strategies tailored to the characteristics of each object. Table 8, in the Appendix, reports how often each of the bodily actions was observed per participant across exhibits. Right-hand anchoring (RA) was observed in all participants, whereas arm movement from front to back (FTB), right-hand anchoring while also contacting the lower part of the exhibit (RAB), and upper-body leaning forward (UF) were rarely observed.

*Phase 1.* Participants freely explored each exhibit and accessed the available information at their own pace. Capture rates varied across exhibits, with the highest proportion of content accessed in Exhibit D (M = 92.8%, SD = 9.5), followed by Exhibit A (M = 73.9%, SD = 15.3), Exhibit C (M = 69.1%, SD = 24.3), and Exhibit B (M = 56.8%, SD = 12.9). The horizontally placed Exhibit D was the most readily accessed, while Exhibit B, a wall-mounted text panel without tactile cues, proved the most challenging. Participants generally found the ring camera interaction easy, with SEQ scores of M =

| Exhibit | Question | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Exhibit A | Q1: It was easy to grasp the overall content of the exhibit using the ring camera. | 6 | 7 | 5 | 6 | 6 | 6 | 7 | 6 | 6 | 6 | 7 |
| | Q2: It was easy to locate a specific section of the exhibit using the ring camera. | 5 | 5 | 6 | 7 | 5 | 6 | 7 | 5 | 6 | 6 | 7 |
| | Q3: It was easy to confirm detailed content of the exhibit using the ring camera. | 5 | 6 | 5 | 7 | 7 | 5 | 6 | 6 | 6 | 5 | 7 |
| Exhibit B | Q4: It was easy to grasp the overall content of the exhibit using the ring camera. | 5 | 5 | 7 | 5 | 6 | 7 | 5 | 6 | 7 | 7 | 7 |
| | Q5: It was easy to locate a specific section of the exhibit using the ring camera. | 2 | 3 | 4 | 2 | 3 | 3 | 5 | 3 | 5 | 5 | 5 |
| | Q6: It was easy to confirm detailed content of the exhibit using the ring camera. | 3 | 4 | 5 | 5 | 4 | 3 | 3 | 6 | 5 | 6 | 5 |
| Exhibit C | Q7: It was easy to grasp the overall content of the exhibit using the ring camera. | 5 | 6 | 6 | 7 | 5 | 7 | 7 | 7 | 7 | 7 | 7 |
| | Q8: It was easy to locate a specific section of the exhibit using the ring camera. | 5 | 5 | 6 | 7 | 7 | 6 | 5 | 5 | 7 | 6 | 7 |
| | Q9: It was easy to confirm detailed content of the exhibit using the ring camera. | 5 | 4 | 5 | 7 | 6 | 7 | 6 | 6 | 7 | 6 | 7 |
| Exhibit D | Q10: It was easy to grasp the overall content of the exhibit using the ring camera. | 6 | 6 | 7 | 6 | 6 | 7 | 6 | 5 | 7 | 6 | 7 |
| | Q11: It was easy to locate a specific section of the exhibit using the ring camera. | 3 | 3 | 5 | 4 | 3 | 3 | 2 | 5 | 3 | 5 | 6 |
| | Q12: It was easy to confirm detailed content of the exhibit using the ring camera. | 4 | 4 | 5 | 4 | 7 | 5 | 5 | 6 | 6 | 7 | 7 |

**Table 2: Results of the three usability questions rated on a 7-point Likert scale (1 = very difficult, 4 = neutral, 7 = very easy), administered after participants completed both Phase 1 (free exploration) and Phase 2 (goal-directed exploration).**

| No. | Item | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 | Median |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Q1 | I was able to understand the relationship between my hand position and the area captured by the camera. | 5 | 6 | 6 | 6 | 6 | 6 | 6 | 7 | 6 | 5 | 7 | 6 |
| Q2 | As the session progressed, I came to understand this relationship more intuitively. | 6 | 6 | 6 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 |
| Q3 | The audio feedback provided by the system matched the level of information I wanted at that moment. | 5 | 5 | 5 | 7 | 5 | 5 | 5 | 6 | 7 | 6 | 6 | 5 |
| Q4 | When trying to grasp the overall content, I had to consciously change the way I moved my hand. | 2 | 3 | 6 | 5 | 6 | 5 | 7 | 2 | 1 | 3 | 3 | 3 |
| Q5 | When searching for specific information, I had to consciously change the way I moved my hand. | 7 | 6 | 6 | 5 | 5 | 7 | 5 | 5 | 7 | 5 | 7 | 6 |
| Q6 | When searching for specific information, I had to consciously adjust how I moved my hand. | 3 | 3 | 3 | 1 | 2 | 2 | 5 | 3 | 2 | 3 | 2 | 3 |

**Table 3: Results of the Likert items responses per participant, after all tasks were completed. Items were rated on a 7-point scale (1 = strongly disagree, 4 = neutral, 7 = strongly agree). For Q4–Q6, higher values indicate a greater need to consciously modify hand movements, implying that natural movements alone were insufficient for the corresponding exploration goal.**

5.73 (SD = 0.90) for Exhibit A, M = 5.82 for Exhibit B, M = 6.45 for Exhibit C, and M = 5.82 for Exhibit D.

*Phase 2.* Participants were allowed to re-hover after each question, and Phase 1 free exploration kept the content fresh in their memory. As a result, most participants were able to answer the questions across three information levels for all four exhibits (Table 7 in

| No. | Question |
|---|---|
| Q1 | How did you develop your understanding of the relationship between your hand position and what the camera was capturing? At what point did it start to feel more intuitive? |
| Q2 | How did you change the way you moved your hand when trying to grasp the overall, section and detail content? |
| Q3 | Were there differences in exploration strategy or difficulty between text-only panels and panels with tactile elements? |
| Q4 | Were there differences between horizontally placed and wall-mounted exhibits? |
| Q5 | How did you feel about exploring exhibits using the ring camera? |
| Q6 | Were there situations where the information returned by the system did not match what you wanted to know? |
| Q7 | What would you like to change or add to the current system? |
| Q8 | Compared with smartphone-based approaches, what are the advantages and disadvantages of the ring camera? |
| Q9 | In what future scenarios do you think the ring camera could be used for reading or exploring information? |

**Table 4: Open-ended questions used in the semi-structured interview.**

| Category | Code | Description |
|---|---|---|
| Anchor hand use | RA | Right hand used as anchor |
| | LA | Left hand used as anchor |
| | B | Both hands in contact with the medium |
| Hovering posture | WH | Hovering the hand on a vertical (wall-mounted) surface |
| | HH | Hovering the hand on a horizontal surface |
| Body posture | C | Crouching (bending the legs) |
| | UB | Upper body leaning backward |
| | UF | Upper body leaning forward |
| Finger action | IU | Lifting only the index finger |
| Extended anchor contact | RAT | Right-hand anchor also contacting the top part of the exhibit |
| | RAB | Right-hand anchor also contacting the bottom part of the exhibit |
| Hand use | R | Only the right hand (no left-hand anchor) |
| Arm and wrist motion | TW | Turning the wrist |
| | BTF | Arm motion from back to front |
| | FTB | Arm motion from front to back |

**Table 5: Coding scheme for bodily actions observed during the user study (15 codes). Actions are grouped by function.**

Appendix). In particular, Exhibit A and C achieved 100% accuracy, and Exhibit B was answered correctly after a few re-hovers. However, for Exhibit D, P1, P2 and P4 could not correctly identify the alternating figure-text layout, in which the five vertically arranged sections (Figure 3 (d)) placed figures on the left and text on the right for even-numbered sections, and the reverse for odd-numbered sections. SEQ scores remained largely stable compared with Phase 1 for Exhibit B (M = 5.55), Exhibit C (M = 6.18), and Exhibit D (M = 5.64), while Exhibit A showed a pronounced improvement (M = 6.73, SD = 0.65), likely reflecting the benefit of the tactile cues once participants had built familiarity with the exhibit.

*Exhibit A.* Across Exhibit A, all participants used the right hand as an anchor (RA) with 174 occurrences. However, different bodily actions emerged for the upper and lower sections. For the upper section, seven participants (P2, P3, P6, P7, P9–P11) hovered their hand in front of the wall (WH) (Figure 4 (a)). To regulate camera distance, two participants (P6, P10) crouched by bending their legs (C), two (P5, P7) leaned their upper body backward (UB), and one (P6) leaned forward (UF). These behaviors show that participants regulated camera coverage using their hands and bodies.

For the lower section placed on a horizontal surface, four participants (P6, P9–P11) hovered their hand above the surface (HH). Two participants (P10, P11) used the left hand in contact with the exhibit as an anchor (LA), while three participants (P5, P9, P10) lifted only the index finger to make fine adjustments to the camera position (IU). When exploring the left-to-right sequence of fetal development, seven participants (P2–P4, P6–P8, P11) placed both hands on the exhibit (B). After capturing information, they touched the models with both hands while listening to the audio descriptions, relating tactile and auditory information.

The ring camera was not used in 49.1% of observations involving the 2.5D fetal model because participants could understand its shape

| Code | Overall | Exhibit A | Exhibit B | Exhibit C | Exhibit D |
|---|---|---|---|---|---|
| RA | 383 | 174 | 73 | 17 | 119 |
| LA | 12 | 12 | 0 | 0 | 0 |
| B | 62 | 55 | 0 | 7 | 0 |
| WH | 55 | 18 | 21 | 16 | 0 |
| HH | 82 | 21 | 2 | 0 | 59 |
| C | 10 | 6 | 0 | 4 | 0 |
| UB | 6 | 4 | 0 | 2 | 0 |
| UF | 1 | 1 | 0 | 0 | 0 |
| IU | 70 | 3 | 2 | 0 | 65 |
| RAT | 17 | 0 | 17 | 0 | 0 |
| RAB | 2 | 0 | 2 | 0 | 0 |
| R | 21 | 21 | 0 | 0 | 0 |
| TW | 3 | 0 | 0 | 3 | 0 |
| BTF | 2 | 0 | 1 | 1 | 0 |
| FTB | 8 | 0 | 8 | 0 | 0 |
| **Total** | **734** | **315** | **126** | **50** | **243** |

**Table 6: Frequencies of bodily action codes across all participants, reported overall and separately for each exhibit.**

through direct touch, suggesting they prioritized tactile exploration when rich tactile information was available.

*Exhibit B.* When accessing sections and detailed content, nine participants (all except P8, P10) used their right hand as an anchor (RA), accounting for 73 occurrences. Of these, 71 occurred while positioning the camera. Participants placed their right hand on the panel to establish their orientation and then moved the camera hand across the panel toward the target area (Figure 4 (b) and (c)).

Four participants (P5, P6, P8, P9) extended this anchor to the upper part of the panel (RAT), with 17 occurrences, whereas only P5 anchored on the lower part (RAB), with two occurrences. This preference suggests that participants generally used the upper edge of the panel as a spatial reference, although the chosen anchor depended on arm reach and comfort.

Participants also adopted strategies beyond right-hand anchoring. Four (P2, P9–P11) hovered their hand in front of the wall-mounted panel (WH), with 21 occurrences, while P8 hovered the hand as if above a horizontal surface (HH) twice. P7 lifted only the index finger bearing the ring camera (IU) twice to make fine camera adjustments when capturing the densely arranged central content. P11 also adjusted camera distance by moving the arm from back to front (BTF) once and from front to back (FTB) eight times.

The translation region on the panel's right side was often abandoned once participants recognized it repeated the same information in another language. Seven participants (P5–P11) deliberately stopped exploring this region. In contrast, the densely arranged central region, containing the laureate's name, the award history and research description, was the most difficult region to capture (Table 2 Q6). Six participants (P1, P2, P4, P6–P8) failed on their initial attempt to capture this region and progressively lowered their arm until the target content entered the camera view.

*Exhibit C.* All participants successfully identified the overall position and accessed the top section. In contrast, only four participants (P2, P6, P10, P11) accessed the bottom section (36.4%). Eight participants (P1–P8) used their right hand as an anchor (RA), with 17 occurrences, while six (P2, P7–P11) hovered their hand in front of the wall surface (WH), with 16 occurrences. These results indicate that participants stabilized the camera either through direct contact with the panel or by hovering in front of the vertical surface.

The large vertical area also required whole-body adjustments. Four participants (P4–P6, P11) crouched once by bending their legs (C) (Figure 5 (a)), while two (P3, P10) leaned their upper body backward (UB) to increase the distance between the camera and the panel (Figure 5 (b)). P6 moved the camera arm from back to front (BTF) once when approaching the panel from a more distant position. Two participants (P1, P10) also rotated their wrist (TW) to adjust the camera orientation (Figure 5 (c)). We revisit this wrist adjustment in Section 4.4. During audio playback, five participants (P2–P4, P7, P8) placed both hands on the panel (B), with seven occurrences. This behavior suggests that after positioning the camera, they returned both hands to the panel and related the audio description to the tactilely distinguishable text and frame.

*Exhibit D.* With participants seated and the catalog placed horizontally, whole body movement was minimal, and participants instead relied on strategic finger and hand movements. During camera positioning, six participants (P1, P4–P6, P9, P11) used their right hand as an anchor (RA), with 119 occurrences. They generally kept their right hand on the page's right edge as a stable reference for the page position and camera–catalog distance, while moving the left hand toward the target content. During audio playback, four participants (P1, P6, P9, P11) maintained their right hand as an anchor, with 49 occurrences.

Five participants (P3, P7–P10) lifted only the index finger bearing the ring camera (IU), with 65 occurrences (Figure 5 (a) and (b)). They kept the other fingers on the catalog and moved the index finger vertically, horizontally, and in depth to access the target text or figure. This enabled fine camera adjustments through finger posture and contact points rather than large hand or body movements. Nine participants (all except P4, P5) hovered their hand above the horizontal surface (HH), with 59 occurrences. P2 and P11 used this strategy throughout the task, whereas others combined hovering with either anchoring or isolated index finger movement.

## 4.2 The Hand as an Anchor to Organize Access

Across all tasks, participants consistently used one hand to place on the medium as a reference point while moving the camera with the other hand. This behavior has been observed in prior work on information access by BLV users [20], yet in our study, it emerged consistently across all media, regardless of tactile cue availability. Although the anchoring strategy served as a single technique, its role varied depending on the medium's characteristics. Two distinct functions of the anchoring hand emerged:

- **Spatial Reference**: The anchor helped participants locate where each element was positioned within the medium, serving as a baseline for constructing the correspondence between the ring camera and retrieved information.


**Figure 4: Participants capturing Exhibits A and B. (a) In Exhibit A, the participant captures the upper section. (b) In Exhibit B, the participant first places the right hand on the upper part of the panel to establish spatial reference before capturing a specific section. (c) The participant then lifts only the right hand while keeping the ring camera aimed at the target.**


**Figure 5: Participants capturing Exhibit C, using not only the arm but the whole body. (a) Crouching down to capture the exhibit. (b) Leaning the upper body backward. (c) Rotating the left wrist to aim the camera at the exhibit.**

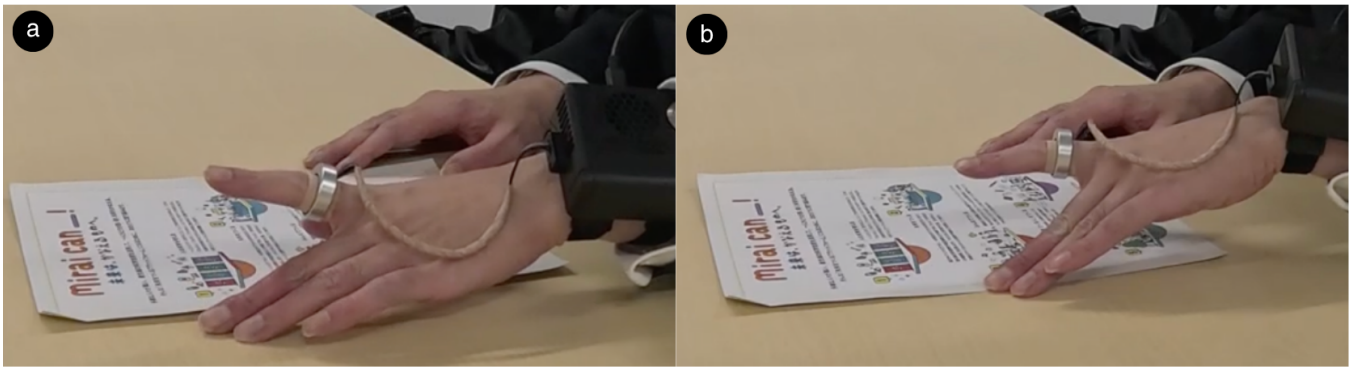


**Figure 6: Participants capturing Exhibit D, where many participants adopted finger-based movement strategies. (a) and (b) The right hand rests on the right side of the catalog to maintain balance, while only the left index finger is lifted and moved to adjust the position of the ring camera.**

- **Postural Support**: The anchor stabilized body balance and camera orientation, particularly for non-tactile media where participants could lose bodily orientation in the air.

*Spatial Reference.* Participants constructed the correspondence between the position of the ring camera and the information it retrieved by using their own hand as a baseline. P3 described this strategy in Exhibit B, explaining that *"Because simply hovering in the air did not give me a sense of where I was, I placed my right hand at the upper front of the panel and then reasoned backward from there to the position I wanted to capture."* Here, P3 used the edge of the medium as an initial baseline and derived the target position relative to the anchoring hand. Similarly, P4 described using the anchor as a marker for what had already been captured, stating that *"When capturing small items, I used my right hand so that I would not capture the same thing multiple times. I kept my right hand on what I had just captured, used that as a distance reference, and then captured what was right below."*

*Postural Support.* For the non-tactile media in particular (Exhibit B and D), eight participants (all except P2–P4) explicitly described the purpose of anchoring not as touching the medium but as stabilizing their own posture and the orientation of the camera. In Exhibit D, P9 explained, *"I kept my right hand on the right edge of the catalog throughout the task. The right hand was not for touching the catalog but only for maintaining the balance of the distance at which I was hovering."* P6 similarly stated, *"In Exhibit B, I have to move my arm in the air, and then I lose a sense of whether I am standing straight. Being able to place one hand on something lets me control the orientation of body."* These accounts indicate that the anchoring function became decoupled from tactile information retrieval and emerged as a modality of its own for securing bodily stability.

*Dynamic Shift Between the Two Functions.* The balance between these two functions shifted according to the characteristics of each medium. In Exhibit A, the anchoring hand acted as a pivot between touch and audio. During the audio explanation, participants engaged both hands with the model while keeping one hand as an anchor to maintain spatial orientation, integrating tactile information with the spoken description. The anchor thus flexibly combined spatial reference and postural support depending on the medium, serving as the central axis that organized participants' information-access behavior. This observation suggests that the ring camera functioned not merely as a text reader, but as an exploratory interface centered on hand movement and body posture.

## 4.3 Adjusting the Camera in Three Dimensions

When accessing information through the ring camera, participants needed to adjust the camera not only in terms of forward and backward distance (Z axis), but also along the vertical and horizontal position on the medium (XY plane). Participants were not informed in advance that different ranges of information would be returned based on distance. Instead, they gradually learned that changing distance affected the scope of the description, although this did not always amount to a clear understanding of all possible levels. This gradual discovery was reflected in the Likert ratings (Table 3, Q1: M = 6, SD = 0.63), and participants reported that the relationship became more intuitive as they moved through the tasks (Table 3, Q2: M = 7, SD = 0.47). Five participants (P1–P3, P8, P11) reported beginning to understand this correspondence with the first medium, another five (P4–P7, P9) with the second medium, and P10 with the third. P5 articulated the three-layer structure that they eventually arrived at, stating that *"I understood that there were three levels of explanation. At the farthest distance, it was the overall picture, in the middle, it was a general description, and when close, I assume it was reading the text."* Participants' bodily adjustments varied across three levels of information access:

- **Overview Capture**: Capturing the overall layout relied on large-scale framing through coordinated arm and torso movements, such as crouching or stepping backward.
- **Section Identification**: Identifying a specific section was the most difficult, as participants lacked both a clear bodily reference for calibration and a means to confirm which range was being captured.
- **Detail Reading**: Reading detailed text relied on precise calibration at the fingertip, isolating the index finger while the other fingers rested on the medium.

*Overview Capture.* When capturing the overall layout of a medium, participants did not simply hold the camera farther away but often used larger-scale framing strategies involving the arm or torso (Table 3, Q4: M = 3, SD = 1.97; lower values indicate that natural movements alone were sufficient). P4 commented *"Just like with the smartphone I normally use, I thought it was better not to change the up-down orientation of the ring camera too much, so I crouched or used my body to adjust the framing, moving my hand and body together."* Similarly, P10 described adjusting stance according to the size of the target when capturing its overall layout, stating *"To capture the whole of an object, because the sizes differ, I would step backward. I adjusted by moving my body to match the size of the object."* P11 reported securing the framing through a shift in the center of gravity for overall-view capture, stating *"For the overall view, I lowered my center of gravity and held my arm far away so that it was aimed at the center of the exhibit."* These accounts show that overview capture often required participants to coordinate hand position with larger bodily framing movements in order to encompass the intended area.

*Section Identification.* Identifying a specific section between the overall overview and detailed reading was the most difficult adjustment (Table 3, Q5: M = 6, SD = 0.94; higher values indicate that natural movements alone were insufficient). P5 stated, *"(The section level) was the hardest. Especially for wall-mounted objects, the sense of distance needs more practice. I think the general-level description is necessary, but being slightly closer than the overall view is intuitively difficult to achieve."* This difficulty did not stem simply from a lack of effort, but from the absence of a clear bodily or perceptual reference for how much closer than an overview view the camera should be held. The difficulty further extended to confirming which range was being captured after the fact. P2 reported mistaking one level for another, stating *"When I thought I was capturing the overall view, I was actually capturing a section. As a result, I got confused about whether the explanation was for the overall view or for a section."* In an extreme case, P8 did not realize that an intermediate level of

description was available in Exhibit B, despite repeatedly moving their arm back and forth while exploring.

*Detail Reading.* When reading detail content participants did not mobilize the whole body but used fine-grained adjustments at the level of the fingertip (Table 3, Q6: M = 3, SD = 1.03; lower values indicate that natural movements alone were sufficient). Participants commonly isolated the index finger bearing the ring camera while keeping the other fingers resting on the medium. P7 described the three-dimensional character of this fingertip adjustment, stating *"I kept all my fingers except the left index finger on the object, and moved the index finger alone, bringing it closer or farther while also shifting it left and right."* P11 similarly articulated the temporal pattern of approach and retreat at the fingertip: *"Basically, for text, my strategy was to bring the (left) index finger closer and then pull it back."* Together, these accounts indicate that detail reading depended on precise calibration of both distance and lateral position at the fingertip, rather than on large-scale bodily repositioning.

## 4.4 Continuous Adjustment of Body Posture

Participants continuously adjusted not only their hand movements, but also their whole-body posture, which became a source of fatigue and burden with certain media. Two characteristics of each medium shaped how postural adjustments unfolded:

- **Placement (Wall-Mounted vs Horizontal)**: Wall-mounted media imposed greater postural burden due to constraints on hand reach and orientation, and participants responded by mobilizing the whole body, while also encountering absolute limits of bodily reach.
- **Tactile Landmarks (Present vs Absent)**: Tactile elements enabled participants to anticipate where to aim the camera, whereas flat media without such cues forced participants to rely on memory and imposed additional cognitive burden.

*Postural Burden with Wall-Mounted Media.* Six participants (P1, P2, P6, P8, P10, P11) reported that horizontal placement was physically less demanding than wall-mounted placement. P2 stated that *"With things on walls, I have to move my whole body away from the object, whereas with horizontal ones I only need to use my arm, so it is easier to understand."* P10 described how constraints on hand orientation directly produced fatigue, reporting that *"With walls, my wrist unconsciously turns upward, and moving my wrist up and down to keep it facing forward was tiring. Horizontal was easier."*

*Alternative Postural Strategies and Their Limits.* In response to such postural burden, participants developed alternative bodily mobilization strategies, while also encountering the absolute limits of their bodily reach. P10 engaged the whole body rather than the hand alone: *When something was on the wall, rather than lowering my hand, I bent my knees or leaned forward. I tried to move not only with my hand but with my whole body."* P1 recounted wrist rotation: *For wall-mounted things, I would normally hold the camera up, but I rotated my wrist around. Since it is difficult to keep my hand up while lowering it, I did this consciously."* P7 switched postural strategies depending on exhibit height: *for low exhibits I had to change the orientation of my hand. The way of moving the hand and the posture needed to change."* P4, in contrast, pointed to the absolute limits of reach: *It was hard to capture objects mounted on walls at low positions, and exhibits above eye level were also difficult,"* describing targets that postural adaptation alone could not compensate for.

*Tactile Landmarks as Reference Points.* With media that offered tactile elements (Exhibit A and C), participants could anticipate where to aim the camera by using those elements as reference points. P4 described how tactile information supported confidence in aiming the camera, stating that *"It is easier to capture things with the ring camera when I can understand them by touch. When I cannot understand anything by touch, I do not know how to touch it. For the fetus, I only need to touch the circular part, and the text is written inside a rectangle, so I can capture with confidence."* P9 reported that tactile landmarks also contributed to calibrating their sense of distance, noting that *"The three-dimensional parts were easier to touch, and my sense of distance gradually came to me."*

*Memory Dependence Without Tactile Cues.* With flat text panels (Exhibit B and D) lacking tactile cues, participants could not rely on physical landmarks and had to rely on memory throughout exploration. P10 verbalized this dependence, stating that *"With things that can be touched, it was easy to understand where each section was and where it was in relation to my body, but without anything to touch, I had to judge from memory, which was difficult."* P11 described the need to restart exploration when memory failed, reporting that *"I thought it was easier to move my arm with things that can be touched. With panels, I have to move the position myself, but if I forget the layout, I have to start over from the beginning."* P3 reported that the absence of tactile cues led to a loss of spatial self-location, stating that *"With a flat panel alone, I did not know how I should look at it. When I moved my hand around, I lost track of where I was. Without cues, searching was hard."* These statements indicate that with flat media, participants had to allocate cognitive resources to maintaining an internal mental map, and cognitive burden emerged in addition to physical burden.

*Weighing Information Value Against Burden.* Many participants felt that the value of acquiring information outweighed physical and cognitive burden. P5 commented, *"I felt, for the first time, that I was reading the text on my own. So this is how sighted people see a layout. Until now, when I took pictures with my smartphone, it would describe the content for me, but I had no sense of reading it myself. It was just a substitute. This time, the fact that I was reading it myself was wonderful."* P11 similarly noted, *"I wanted to read the information, so I did not think anything about my arm getting tired or having to move it."* However, P2 noted: *"Certainly, (the ring camera) let me understand the layout, and I did not need to hold it up for the translation part because that was unnecessary, but I did not feel the need to go out of my way to use my arm to read text that I could not touch,"* suggesting that this trade-off depended on both media characteristics and individual differences.

## 4.5 The System's Provision of Information

Participants generally viewed the ability to obtain information at different ranges by varying distance as useful. P5 connected close-up captures with tactile exploration at hand, stating that *Close-up captures are easy to understand. When my hand is close to the object, I can easily imagine that text is written here."* P7 valued the ability to judge early which information to omit: *When I captured*

*a section, the system told me 'this section is translated into English.' Since English was fine, I deliberately skipped it. When I normally use smartphone apps, they describe everything that is visible. Most of the time I feel that the information is unnecessary, so it was helpful that the system provided different amounts of information."* P9, with low vision, noted that distance-dependent explanations helped confirm ambiguous information obtained through residual vision: *"When I held my hand in the middle area, the system started describing the section. When I got closer, it started describing the details. What I had been seeing vaguely became clearer by changing the distance."*

While participants valued the multi-granularity presentation, they identified challenges in how the system provided information:

- **Level Identification**: Participants found it difficult to immediately identify which granularity (overview, section, or detail) was currently being provided, and proposed additional cues such as distinct sounds or heading-like announcements.
- **Hand-Audio Correspondence**: The audio did not always align with hand position, as the reading began from the start of a text block regardless of where participants were hovering within the region.
- **Capture State Feedback**: The system lacked feedback on capture state, such as when the target was missing, when a blank area was captured, when another exhibit appeared in the frame, or when participants were too close.
- **Control Granularity**: Participants differed in the type and granularity of control they sought, from navigation buttons for adjacent items to modality-switching based on whether the medium was tactile.

*Level Identification.* The multi-granularity presentation made it difficult to immediately identify which range of information was currently being provided. As shown in Section 4.3, participants found intermediate distances particularly difficult, and this translated into difficulty judging whether the returned explanation corresponded to overview, section, or detail. Participants proposed additional cues. P1 stated, *"It would be good if different sounds were played for overview, section, and detail when holding up the camera succeeded."* P9 said, *"It might be good if it explained like a screen reader saying 'Heading 1' or 'Heading 2.' It does not need to say the content. Just telling me where what is written would be enough."* This suggests the need for meta-information that signals the hierarchy level of the current explanation before the content is read.

*Hand-Audio Correspondence.* What participants sought was not merely switching between granularities. Even when the intended level of explanation was returned, it did not always align with hand position, making it difficult to relate audio content to the hand. In the current prototype, when a text region is detected, the explanation begins from the start of the detected text block regardless of where within the region the participant is hovering. As a result, even when a participant was hovering around the center, the reading did not necessarily begin from the text corresponding to the center. P6 stated, *"When I held my hand at the upper right and then around the middle, the positions were clearly different, but the same explanation was played. It would be better if what is written differed according to the position of my hand."*

*Capture State Feedback.* The absence of immediate feedback on camera orientation and capture state was also a common challenge. The system did not notify participants when the target was not captured, when a blank area was being captured, when another exhibit appeared in the frame, or when the participant was too close. P4 stated, *"It would be good if the system told me, when I captured something, things like 'another exhibit is in the frame,' 'you are too close,' or 'there is nothing here."'* P7 said, *"Feedback is needed for areas where nothing is captured. I do not want to waste time there."* P6, referring to an existing similar system, stated, *"Seeing AI says things like 'the upper right is not in frame' or 'stay there,' so if the system could notice that there was nothing and then give instructions like 'move away,' that would be good,"* asking not merely for a notification of capture success or failure, but for instructive feedback that indicated how to move the hand or camera.

*Control Granularity.* Participants differed in the content and granularity of control they sought. P3 stated, *"Suppose I read what was written in the middle, for example. When I think that something might be written before or after it, it would be good if pressing a button read the previous item, and pressing it twice read the next item,"* seeking navigation controls that moved to information before or after the current item based on what was being read. P9, on the other hand, stated, *"I think feedback is good for things that can be touched, and when you cannot touch something, just reading aloud is enough. I thought it would be smoother if the delivery method changed between tactile and non-tactile things,"* pointing out that the type of feedback needed changed depending on whether tactile cues were available.

## 5 System Design Considerations

From the behaviors and challenges observed in Section 4, we derive four design goals (D1–D4) for future ring-camera-based systems.

### 5.1 D1: Supporting One-Handed Exploration Grounded in Anchor Use

Our previous work showed that BLV users frequently maintained contact with an object using one hand while exploring with the ring camera hand [58]. In this study, however, we found that the non-wearing hand acted not only as a spatial reference to understand the medium's layout but also to maintain postural stability while the camera hand moved in mid-air (Section 4.2). This bodily strategy emerged across all participants without explicit system support.

To naturally incorporate such embodied practice, future systems need to be designed under the premise that an anchor hand can be present. Specifically, by interpreting the motion of the ring camera hand in relation to the anchor hand, the system can more accurately infer users' exploratory intent. Prior research has already indicated that hand-object interactions improve recognition accuracy in egocentric vision [30, 32], and our findings extend this line of work to the information exploration tasks of BLV users.

In our study, the distinction between the anchor hand and the camera hand was not implemented, and researchers interpreted the roles of the two hands through a WoZ setup. Due to this limitation, the system sometimes failed to accurately capture the user's intent when participants moved their anchor hand or grasped the target with both hands (i.e., when a hand entered the frame, the system did not provide audio). Similar challenges have been reported for

wearable visual assistance systems in general [9], and the mapping between the bodily actions of BLV users and the delivery of information remains an open issue.

In future ring camera-based system implementations, the anchor hand can be detected within the wide-angle camera view and used as a reference point for interpreting movement. When the anchor hand remains stationary, it serves as a fixed reference point, allowing the motion of the camera hand (e.g., scanning, or approaching) to be interpreted relative to it. This enables the system to explicitly support the strategy of "reasoning backward from the anchor" (P4) that participants developed on their own.

## 5.2 D2: Making Scale Hierarchy Explicit and Confirmable

BLV users discovered that varying distance yielded different information granularities, but this understanding was not consistently stable—particularly at the intermediate granularity for non-tactile objects, where users lacked a means to confirm whether information was presented at all (Section 4.3). Mechanisms for conveying hierarchical structure through audio are already standardized in web accessibility through screen reader output [59]. However, excessive reading can cause information overload [62], and when information is presented under system initiative, a gradual progression from overall to finer levels is recommended [57].

By contrast, our WoZ setup adopted a user-driven design in which body motion determined the information obtained. Although participants could access a wide range of information through body movement, the system began reading only when the medium was captured at the correct distance. Consequently, users could not immediately confirm whether they had reached the intended granularity or whether silence meant "capture failed" or "no appropriate information available." In future systems, hierarchical information could be conveyed before reading by assigning different timbres and rhythms to each level [8]. The sound-to-level mapping should be determined through participatory design with BLV users [38].

Beyond switching between granularities, correspondence with position within each granularity is also important. In our system, reading began at the start of each detected text block, regardless of the participant's hovering position. Future systems could use the camera's field of view to determine where reading begins within the text block. Combined with the anchor-relative position estimation described in D1 (Section 5.1), this design reflects the user's exploratory intent of "where they are trying to read now." Prior work shows that BLV users' information needs vary by region and context within an image [54], and reflecting hand position, a bodily input, in the reading start point can be understood as a bodily instantiation of this diversity.

## 5.3 D3: Providing Real-Time Feedback on Capture State

Participants strongly requested a means to immediately confirm what they were capturing with the camera at any given moment (Section 4.5). In the current system, there was no mechanism to notify participants in real time when no potential target was captured. Prior work on photography support for BLV users has consistently pointed out the importance of real-time notification of capture success and the state of the target [5, 28]. Commercial applications such as SeeingAI [35] also provide features that verbally notify users of the direction and proximity of targets outside the field of view. Prior work has further shown that instructive feedback, which indicates how to move rather than simply notifying of failure, is effective [9, 26], and future systems need to follow this principle.

In future systems, Optical Character Recognition (OCR) could be used to detect text regions, with areas where no text is detected treated as candidate "blank" regions. However, OCR may fail under conditions such as small text or poor lighting [42]. To address this, Vision-Language Models (VLMs) could be used alongside OCR to distinguish between missed text and genuinely blank regions. Detection of out-of-frame content, the presence of other objects, and excessive proximity could be achieved by combining VLM outputs with IMU data, as well as monocular depth estimation and bounding box size (as described in D2). VLMs can run via cloud APIs or on-device, with local models preferable for privacy-sensitive tasks [21, 51]. Additionally, environments like science museums often include domain-specific terminology and exhibits that general-purpose VLMs may not recognize reliably [58]. In such cases, task-specific models (e.g., trained with YOLO [45]) could improve recognition performance.

## 5.4 D4: User-Driven Control and Cognitive Offloading

In our study, information delivery progressed automatically in response to the participant's body motion, and the system provided almost no active controls on the user's side, such as stopping, repeating, or adjusting the speed of the audio. Moreover, because the system did not retain exploration history, participants had to rely entirely on their own memory to distinguish regions they had already explored from those they had not, and to reconstruct the layout when they lost track mid-exploration.

Reducing reliance on memory is known as cognitive offloading [47], in which external tools take on information that users would otherwise have to hold internally. In systems for BLV users, the retention of exploration history and the externalization of spatial information have similarly been discussed as design approaches that reduce the burden of mental map construction [40].

Building on these insights, future systems could provide physical buttons on the ring camera for navigation operations such as moving to the previous or next region and re-reading the current one, as well as LLM-based features for summarization requests and free-form questions. These mechanisms would allow participants to actively adjust the flow of information according to their own pace and interest, rather than receiving it passively. Such customization is consistent with prior findings on the importance of adapting guidance to individual BLV users [19, 57].

# 6 Discussion and Future Work

We draw lessons learned and design implications for future ring-camera-based systems supporting BLV users.

## 6.1 Hand-Mediated Inquiry: Reframing the Ring Camera as an Exploratory Interface

Our observations show that the ring camera functioned not simply as a text-to-speech device, but as an exploratory interface centered on hand movement and body posture. Across all media, participants placed one hand as an anchor. This mode of engagement is qualitatively different from the passive reception of information found in conventional audio guides or OCR applications. In the interaction with the ring camera, information was retrieved through the user's bodily motion, and the user actively composed the information space through the position and movement of their own body.

Such active engagement resonates with prior work showing that reading is, by its nature, an act of retrieving information through one's own bodily motion rather than a passive reception [14]. The "feeling, for the first time, that I was reading on my own" expressed by P5 in Section 4.4 is structurally parallel to this insight.

At the same time, the value of bodily active engagement was not uniform across participants. P2 stated that *"I did not feel the need to go out of my way to use my arm to read text that I could not touch,"* indicating that the balance between physical effort and informational value depends on individual visual ability and information needs. This individual variation connects to the principles of ability-based design [60], which emphasize adaptation to the abilities and preferences of individual users. Future systems need to be designed so that users can adjust the depth of their bodily engagement according to their own preferences, rather than being required to engage uniformly.

## 6.2 The Asymmetry of Burden Between Tactile and Non-Tactile Media

In our study, differences in exploration strategies between media with and without tactile landmarks were not merely about usability; they produced qualitative differences in how BLV users accessed information. With media offering tactile elements (Exhibit A), participants could anticipate the scope of exploration by using those elements as reference points and establish the correspondence between hand and information through the medium itself. By contrast, with flat media lacking tactile cues (Exhibits B and D), participants had to maintain this correspondence through memory and bodily orientation (Section 4.4).

This difference can be understood through reference frames, a central concept in haptic cognition research. Prior work shows that BLV users' haptic exploration relies on both external and body-centered reference frames, with greater dependence on the latter when tactile landmarks are unavailable [36]. When tactile landmarks provided an external reference frame, spatial structure was externalized onto the medium, freeing cognitive resources to understand information. Without such landmarks, participants performed a dual task: maintaining spatial structure internally while processing information.

This asymmetry reframes support for non-tactile media in BLV assistive systems. The difficulty of non-tactile media arises not from the absence of tactile cues themselves, but from the lack of an external reference frame; support should therefore provide such a frame through another modality. D1 (Section 5.1) and D3 (Section 5.3) instantiate this direction.

## 6.3 Differences in Strategies Between Blind and Low Vision Users

Our study included both six blind and five low vision participants, and we observed qualitative differences in how they used the ring camera. Low vision participants used the ring camera to clarify and verify partially perceptible visual information, with audio descriptions reinforcing what they perceived through residual vision. When partial reading was possible visually, however, some chose to avoid the physical burden of arm-based exploration, adjusting the depth of engagement based on residual vision and the characteristics of the medium.

For blind participants, in contrast, the ring camera was the primary setting for information exploration rather than a supplementary tool. They developed multi-stage bodily strategies that combined tactile landmarks, the anchor hand, and bodily orientation, flexibly shifting between them according to the medium. For some, this exploration held meaning beyond convenience, offering an autonomous exploring experience unavailable through prior speech-based substitutes.

These differences suggest that future systems should offer interaction modes adapted to visual ability: rich multimodal feedback with anchor-hand and auditory-landmark support (D1, D3) for blind users, and concise complementary information that minimizes physical burden for low vision users. Such adaptation aligns with ability-based design [60], which emphasizes responsiveness to abilities rather than treating BLV users as a single category.

## 6.4 Limitations

Our WoZ setup approximated distance judgment, information hierarchy selection, and text recognition rather than implementing them automatically. Thus, our findings do not directly evaluate D1–D4 in practice, and their feasibility must be validated in implemented systems. It may also have underestimated behavioral adjustments to delayed, incorrect, or missing automated outputs. Recognition latency and false negatives may cause users to pause, repeat movements, or alter exploration strategies. The camera's 0.7 MP resolution may also limit real-world recognition, particularly of small or distant content, potentially requiring higher-resolution cameras. Future systems should communicate processing status and uncertainty through audio or haptic cues for ongoing processing, low confidence, or recognition failure.

All materials came from a museum setting and represented diverse combinations of textual and tactile information media. Rather than isolating factors such as tactile features, font size, content density, or spatial orientation, we examined how BLV users adapted their exploration strategies across realistic configurations. Because these variables were bundled, exhibit differences cannot be attributed to any single factor. Although this diversity revealed bodily strategies across varied media, controlled studies should systematically vary individual properties to identify which features shape these behaviors. Museums do not represent the full range of everyday media, such as documents, labels, and public signage, so studies in other contexts are needed to assess generalizability.

Some strategies may have reflected the three-level audio hierarchy rather than the exhibits' physical characteristics alone. Hand anchoring and postural adjustment appeared to arise mainly from

the physical demands of maintaining contact, orientation, and balance, whereas moving closer for details or farther for an overview may partly reflect the imposed correspondence between distance and information granularity. Future work should therefore examine their applicability to systems with different information structures.

## 7 Conclusion

In this paper, we investigated the spatial interaction strategies that BLV users employ when accessing information through a ring camera. Through a WoZ study with eleven BLV participants exploring four exhibits that varied in the availability of tactile cues, we characterized a consistent bodily pattern: participants used one hand as an anchor while adjusting the camera-wearing hand three-dimensionally, with retrieval spanning from torso-level framing to fingertip-level calibration, and with intermediate-range capture remaining particularly challenging.

The anchoring hand served as both a spatial reference for locating information and a postural support for stabilizing balance and camera orientation. It functioned as an organizing axis rather than passive support. This pattern appeared across all media, but its success depended on tactile cues. Without them, participants had to maintain the spatial structure in memory.

Building on these observations, we derived four design goals for ring camera-based systems: supporting one-handed exploration grounded in anchor use, making scale hierarchy explicit and confirmable, providing real-time feedback on capture state, and enabling user-driven control with cognitive offloading. Our findings reframe the ring camera from a text-reading device to an exploratory interface centered on hand–body coordination, pointing toward embodied information access for BLV users across media.

## Acknowledgments

We sincerely thank Sony Semiconductor Solutions Corporation for their contributions to the hardware design and implementation. This work was supported by JSPS KAKENHI (JP25KJ0653). João Guerreiro was partially funded by FCT – Fundação para a Ciência e a Tecnologia, through the LASIGE Research Unit, DOI: https://doi.org/10.54499/UID/00408/2025, and the FCT Mobility program, ref. FCT/Mobility/1379258621/2024-25.

## References

[1] Vikas Ashok, Yevgen Borodin, Svetlana Stoyanchev, Yuri Puzis, and I. V. Ramakrishnan. 2014. Wizard-of-Oz evaluation of speech-driven web browsing interface for people with vision impairments. In *Proceedings of the 11th Web for All Conference* (Seoul, Korea) *(W4A '14)*. Association for Computing Machinery, New York, NY, USA, Article 12, 9 pages. doi:10.1145/2596695.2596699

[2] Elisabeth Salzhauer Axel and Nina Sobol Levent. 2003. *Art beyond sight: a resource guide to art, creativity, and visual impairment.* American Foundation for the Blind.

[3] Be My Eyes. 2026. Be My AI. https://www.bemyeyes.com/bme-ai/. Accessed: 2026-04-01.

[4] Hugh Beyer and Karen Holtzblatt. 1999. Contextual design. *interactions* 6, 1 (1999), 32–42.

[5] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, 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*. 333–342.

[6] Roger Boldu, Alexandru Dancu, Denys J.C. Matthies, Thisum Buddhika, Shamane Siriwardhana, and Suranga Nanayakkara. 2018. FingerReader2.0: Designing and Evaluating a Wearable Finger-Worn Camera to Assist People with Visual Impairments while Shopping. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 94 (Sept. 2018), 19 pages. doi:10.1145/3264904

[7] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology* 18, 3 (2021), 328–352. arXiv:https://doi.org/10.1080/14780887.2020.1769238 doi:10.1080/14780887.2020.1769238

[8] Stephen Brewster, Veli-Pekka Raty, and Atte Kortekangas. 1996. Earcons as a method of providing navigational cues in a menu hierarchy. In *People and Computers XI: Proceedings of HCI'96*. Springer, 169–183.

[9] Ruei-Che Chang, Yuxuan Liu, and Anhong Guo. 2024. Worldscribe: Towards context-aware live visual descriptions. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–18.

[10] Michael Cutter and Roberto Manduchi. 2017. Improving the accessibility of mobile OCR apps via interactive modalities. *ACM Transactions on Accessible Computing (TACCESS)* 10, 4 (2017), 1–27.

[11] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies: why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces*. 193–200.

[12] Josh Urban Davis, Te-Yen Wu, Bo Shi, Hanyi Lu, Athina Panotopoulou, Emily Whiting, and Xing-Dong Yang. 2020. TangibleCircuits: An Interactive 3D Printed Circuit Education Tool for People with Visual Impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376513

[13] Steven Dow, Blair MacIntyre, Jaemin Lee, Christopher Oezbek, Jay David Bolter, and Maribeth Gandy. 2005. Wizard of Oz Support throughout an Iterative Design Process. *IEEE Pervasive Computing* 4, 4 (Oct. 2005), 18–26. doi:10.1109/MPRV.2005.93

[14] Yasmine N. El-Glaly, Francis Quek, Tonya Smith-Jackson, and Gurjot Dhillon. 2013. Touch-screens are not tangible: fusing tangible interaction with touch glass in readers for the blind. In *Proceedings of the 7th International Conference on Tangible, Embedded and Embodied Interaction* (Barcelona, Spain) *(TEI '13)*. Association for Computing Machinery, New York, NY, USA, 245–253. doi:10.1145/2460625.2460665

[15] Envision. 2026. Envision: Speaks Visual Information Aloud. https://www.letsenvision.com/. Accessed: 2026-04-23.

[16] Isabela Figueira, Yoonha Cha, and Stacy M. Branham. 2024. Intersecting Liminality: Acquiring a Smartphone as a Blind or Low Vision Older Adult. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 45, 14 pages. doi:10.1145/3663548.3675622

[17] Leah Findlater, Lee Stearns, Ruofei Du, Uran Oh, David Ross, Rama Chellappa, and Jon Froehlich. 2015. Supporting everyday activities for persons with visual impairments through computer vision-augmented touch. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*. 383–384.

[18] James J Gibson. 1962. Observations on active touch. *Psychological review* 69, 6 (1962), 477.

[19] Ricardo E Gonzalez Penuela, Jazmin Collins, Cynthia Bennett, and Shiri Azenkot. 2024. Investigating Use Cases of AI-Powered Scene Description Applications for Blind and Low Vision People. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 901, 21 pages. doi:10.1145/3613904.3642211

[20] Tiago Guerreiro, Kyle Montague, João Guerreiro, Rafael Nunes, Hugo Nicolau, and Daniel JV Gonçalves. 2015. Blind people interacting with large touch surfaces: Strategies for one-handed and two-handed exploration. In *Proceedings of the 2015 International Conference on Interactive Tabletops & Surfaces*. 25–34.

[21] Yean Li Ho, Siong Hoe Lau, and Afizan Azman. 2024. Picture superiority effect in authentication systems for the blind and visually impaired on a smartphone platform. *Universal Access in the Information Society* 23, 1 (2024), 179–189.

[22] Leona Holloway, Matthew Butler, and Kim Marriott. 2023. TactIcons: Designing 3D Printed Map Icons for People who are Blind or have Low Vision. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 543, 18 pages. doi:10.1145/3544548.3581359

[23] Leona Holloway, Matthew Butler, Alex Waddell, and Kim Marriott. 2025. 3D Printing for Accessible Education: A Case Study in Assistive Technology Adoption. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 273, 19 pages. doi:10.1145/3706598.3713689

[24] Leona Holloway, Kim Marriott, and Matthew Butler. 2018. Accessible Maps for the Blind: Comparing 3D Printed Models with Tactile Graphics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3173574.3173772

[25] Crescentia Jung, Jazmin Collins, Ricardo E. Gonzalez Penuela, Jonathan Isaac Segal, Andrea Stevenson Won, and Shiri Azenkot. 2024. Accessible Nonverbal Cues

to Support Conversations in VR for Blind and Low Vision People. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 20, 13 pages. doi:10.1145/3663548.3675663

[26] Hernisa Kacorri, Kris M Kitani, Jeffrey P Bigham, and Chieko Asakawa. 2017. People with visual impairment training personal object recognizers: Feasibility and challenges. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 5839–5849.

[27] Susan J Lederman and Roberta L Klatzky. 1987. Hand movements: A window into haptic object recognition. *Cognitive psychology* 19, 3 (1987), 342–368.

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

[29] Kyungjun Lee, Jonggi Hong, Ebrima Jarjue, Ernest Essuah Mensah, and Hernisa Kacorri. 2022. From the lab to people's home: lessons from accessing blind participants' interactions via smart glasses in remote studies. In *Proceedings of the 19th international web for all conference*. 1–11.

[30] Kyungjun Lee and Hernisa Kacorri. 2019. Hands holding clues for object recognition in teachable machines. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[31] Kyungjun Lee, Daisuke Sato, Saki Asakawa, Hernisa Kacorri, and Chieko Asakawa. 2020. Pedestrian detection with wearable cameras for the blind: A two-way perspective. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[32] Kyungjun Lee, Abhinav Shrivastava, and Hernisa Kacorri. 2023. Leveraging Hand-Object Interactions in Assistive Egocentric Vision. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 6 (June 2023), 6820–6831. doi:10.1109/TPAMI.2021.3123303

[33] Nina Levent, Joan Pursley, and Leigh Ann Mesiti. 2011. Speaking Out on Art and Museums. (2011).

[34] Jingyi Li, Son Kim, Joshua A. Miele, Maneesh Agrawala, and Sean Follmer. 2019. Editing Spatial Layouts through Tactile Templates for People with Visual Impairments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3290605.3300436

[35] Microsoft. 2026. Seeing AI: A Visual Assistant for the Blind. https://www.seeingai.com/. Accessed: 2026-04-01.

[36] Susanna Millar. 2008. *Space and sense*. Psychology Press.

[37] Irene Miller, Aquinas Pather, Janet Milbury, Lucia Hasty, Allison O'Day, Diane Spence, and S Osterhaus. 2010. Guidelines and standards for tactile graphics. *The Braille Authority of North America* (2010).

[38] Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Commun. ACM* 36, 6 (1993), 24–28.

[39] Ruth G Nagassa, Matthew Butler, Leona Holloway, Cagatay Goncu, and Kim Marriott. 2023. 3D Building Plans: Supporting Navigation by People who are Blind or have Low Vision in Multi-Storey Buildings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 539, 19 pages. doi:10.1145/3544548.3581389

[40] Vishnu Nair, Hanxiu'Hazel' Zhu, and Brian A Smith. 2023. ImageAssist: tools for enhancing touchscreen-based image exploration systems for blind and low vision users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[41] Suranga Nanayakkara, Roy Shilkrot, Kian Peen Yeo, and Pattie Maes. 2013. EyeRing: a finger-worn input device for seamless interactions with our surroundings. In *Proceedings of the 4th Augmented Human International Conference* (Stuttgart, Germany) *(AH '13)*. Association for Computing Machinery, New York, NY, USA, 13–20. doi:10.1145/2459236.2459240

[42] Francisca O Nwokoma, Juliet N Odii, Ikechukwu I Ayogu, and James C Ogbonna. 2021. Camera-based OCR scene text detection issues: A review. *World Journal of Advanced Research and Reviews* 12, 3 (2021), 484–489.

[43] OrCam. 2026. OrCam: AI-Driven Assistive Technology for the Blind and Visually Impaired. https://www.orcam.com/en-us/home. Accessed: 2026-04-23.

[44] Joshua Rader, Troy McDaniel, Artemio Ramirez Jr, Shantanu Bala, and Sethuraman Panchanathan. 2014. A Wizard of Oz Study Exploring How Agreement/Disagreement Nonverbal Cues Enhance Social Interactions for Individuals Who Are Blind. In *International Conference on Human-Computer Interaction*. Springer, 243–248.

[45] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[46] Samuel Reinders, Swamy Ananthanarayan, Matthew Butler, and Kim Marriott. 2023. Designing Conversational Multimodal 3D Printed Models with People who are Blind. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 2172–2188. doi:10.1145/3563657.3595989

[47] Evan F Risko and Sam J Gilbert. 2016. Cognitive offloading. *Trends in cognitive sciences* 20, 9 (2016), 676–688.

[48] Jeff Sauro and Joseph S. Dumas. 2009. Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 1599–1608. doi:10.1145/1518701.1518946

[49] Suraj Singh Senjam. 2021. Smartphones as assistive technology for visual impairment. *Eye* 35, 8 (2021), 2078–2080.

[50] Anchal Sharma, PV Madhusudhan Rao, and Srinivasan Venkataraman. 2022. Object recognition from two-dimensional tactile graphics: What factors lead to successful identification through touch?. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–5.

[51] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615* (2025).

[52] Roy Shilkrot, Jochen Huber, Wong Meng Ee, Pattie Maes, and Suranga Chandima Nanayakkara. 2015. FingerReader: a wearable device to explore printed text on the go. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2363–2372.

[53] Roy Shilkrot, Jochen Huber, Jürgen Steimle, Suranga Nanayakkara, and Pattie Maes. 2015. Digital digits: A comprehensive survey of finger augmentation devices. *ACM Computing Surveys (CSUR)* 48, 2 (2015), 1–29.

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

[55] Lee Stearns, Ruofei Du, Uran Oh, Catherine Jou, Leah Findlater, David A Ross, and Jon E Froehlich. 2016. Evaluating haptic and auditory directional guidance to assist blind people in reading printed text using finger-mounted cameras. *ACM Transactions on Accessible Computing (TACCESS)* 9, 1 (2016), 1–38.

[56] Shan-Yuan Teng, Gene SH Kim, Xuanyou Liu, and Pedro Lopes. 2025. Seeing with the hands: A sensory substitution that supports manual interactions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–15.

[57] Ayaka Tsutsui, Xiyue Wang, Hironobu Takagi, and Chieko Asakawa. 2025. Investigating "Touch and Talk" for Blind and Low Vision People: Science Communication Assistance Through Exploring Multiple Tactile Objects. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–21.

[58] Ayaka Tsutsui, Xiyue Wang, Hironobu Takagi, Yoichi Ochiai, and Chieko Asakawa. 2026. Eyes on the Palm: Investigating a Ring-Shaped Camera for Seamless Accessible Tactile Exploration. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–24.

[59] WebAIM. [n. d.]. Introduction to ARIA - Accessible Rich Internet Applications. https://webaim.org/techniques/aria/. Accessed: 2026-04-22.

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

[61] Yoshiyuki Koyanagi. 2026. Swift Eyes. https://apps.apple.com/jp/app/swift-eyes/id6742831929. Accessed: 2026-04-01.

[62] Yaman Yu, Bektur Ryskeldiev, Ayaka Tsutsui, Matthew Gillingham, and Yang Wang. 2025. From cluttered to clear: Improving the web accessibility design for screen reader users in e-commerce with generative AI. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–19.

[63] Kaixing Zhao, Sandra Bardot, Marcos Serrano, Mathieu Simonnet, Bernard Oriola, and Christophe Jouffrais. 2021. Tactile fixations: A behavioral marker on how people with visual impairments explore raised-line graphics. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

# A Appendix

| Exhibit | Question Type | Question |
|---|---|---|
| Exhibit A | Overview | What is this exhibit about, and how many sub-displays does it consist of? |
| | Section | How many weeks old is the fetus shown in the front display? |
| | Detail | How large is the fetus at 50 days of development? |
| Exhibit B | Overview | What was the content of the exhibit, arranged from left to right? |
| | Section | Where on the panel is this person's research described? |
| | Detail | In what year and for what category did this person receive the award? |
| Exhibit C | Overview | In which languages is the exhibit written? |
| | Section | What is this exhibit about? |
| | Detail | What do cells need to produce in order for the body to function properly? |
| Exhibit D | Overview | How is the content of this page arranged? |
| | Section | Where on the page is the section about "Building Together" located? |
| | Detail | The section "Thinking about the Future of People" includes four themes; what are they? |

**Table 7: Three questions posed to participants for each of the four exhibits, categorized by question type.**

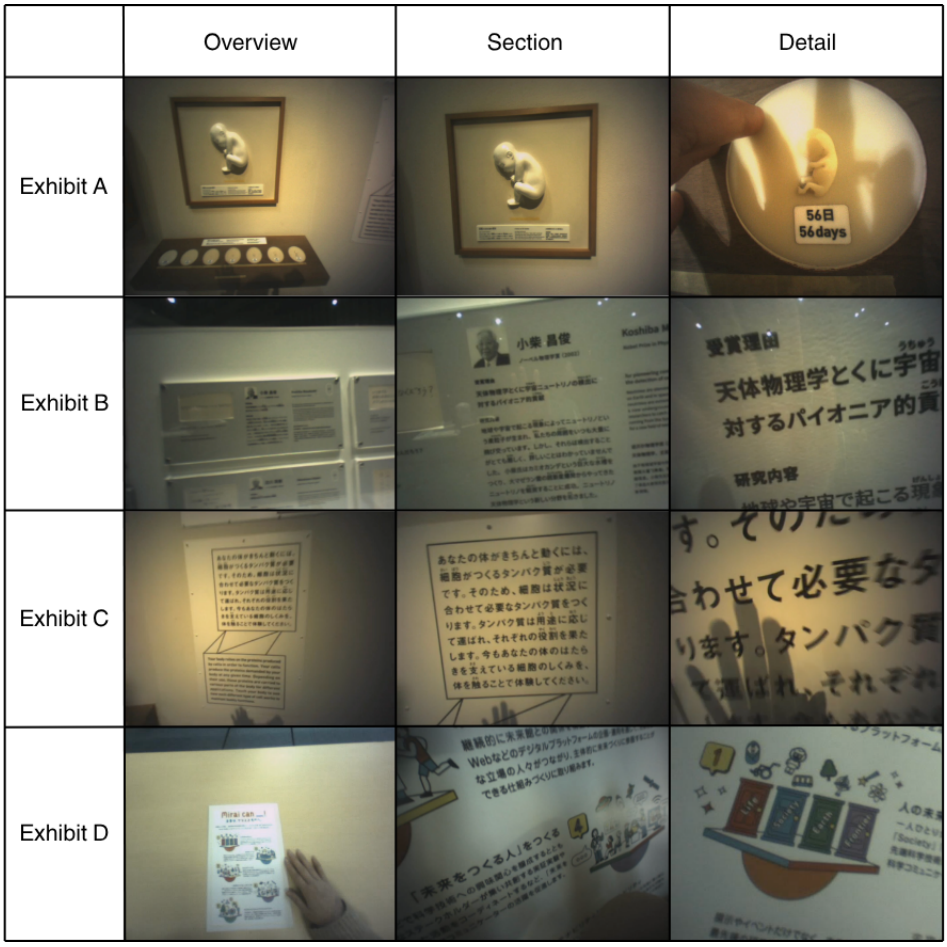


**Figure 7: Examples of ring camera images at each information level across the four exhibits. Each row corresponds to Exhibits A–D, and the columns show an overview of the entire exhibit, a section within the exhibit, and a detail view of specific text or an object element.**

| | *Each entry: Exhibit A/B/C/D* | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | RA | LA | B | WH | HH | C | UB | UF | IU | RAT | RAB | R | TW | BTF | FTB | Overall |
| P1 | 48/15/1/34 | – | – | – | 0/0/0/1 | – | – | – | – | – | – | – | 0/0/2/0 | – | – | 101 |
| P2 | 11/5/1/0 | – | 7/0/1/0 | 2/1/3/0 | 0/0/0/19 | – | – | – | – | – | – | – | – | – | – | 50 |
| P3 | 16/13/2/0 | – | 7/0/2/0 | 2/0/0/0 | 0/0/0/1 | – | 0/0/1/0 | – | 0/0/0/18 | – | – | – | – | – | – | 62 |
| P4 | 18/13/3/19 | – | 11/0/2/0 | – | – | 0/0/1/0 | – | – | – | – | – | – | – | – | – | 67 |
| P5 | 19/5/4/16 | – | – | – | – | 0/0/1/0 | 3/0/0/0 | – | 1/0/0/0 | 0/6/0/0 | 0/2/0/0 | 13/0/0/0 | – | – | – | 70 |
| P6 | 8/8/4/36 | – | 15/0/0/0 | 1/0/0/0 | 1/0/0/1 | 3/0/1/0 | – | 1/0/0/0 | – | 0/1/0/0 | – | 1/0/0/0 | – | 0/0/1/0 | – | 82 |
| P7 | 14/8/1/0 | – | 7/0/1/0 | 2/0/1/0 | 0/0/0/1 | – | 1/0/0/0 | – | 0/2/0/15 | – | – | – | – | – | – | 53 |
| P8 | 17/0/1/0 | – | 7/0/1/0 | 0/0/1/0 | 0/2/0/6 | – | – | – | 0/0/0/10 | 0/7/0/0 | – | – | – | – | – | 52 |
| P9 | 14/4/0/13 | – | – | 3/5/3/0 | 1/0/0/5 | – | – | – | 1/0/0/9 | 0/3/0/0 | – | 7/0/0/0 | – | – | – | 68 |
| P10 | 1/0/0/0 | 7/0/0/0 | – | 1/12/4/0 | 7/0/0/6 | 3/0/0/0 | 0/0/1/0 | – | 1/0/0/13 | – | – | – | 0/0/1/0 | – | – | 57 |
| P11 | 8/2/0/1 | 5/0/0/0 | 1/0/0/0 | 7/3/4/0 | 12/0/0/19 | 0/0/1/0 | – | – | – | – | – | – | – | 0/1/0/0 | 0/8/0/0 | 72 |

**Table 8: Frequencies of bodily action codes by participant and exhibit. Each cell reports the number of occurrences in the order Exhibit A, Exhibit B, Exhibit C, and Exhibit D, separated by slashes. Camera positioning and audio playback are combined. A dash indicates that the action was not observed for that participant in any of the four exhibits. The Overall column reports the total number of coded action occurrences for each participant across all action codes and exhibits.**