TouchScribe: Augmenting Non-Visual Hand-Object Interactions with Automated Live Visual Descriptions

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

Rosiana Natalie
Michigan Institute for Data & AI in Society
University of Michigan
Ann Arbor, Michigan, USA
rosianan@umich.edu

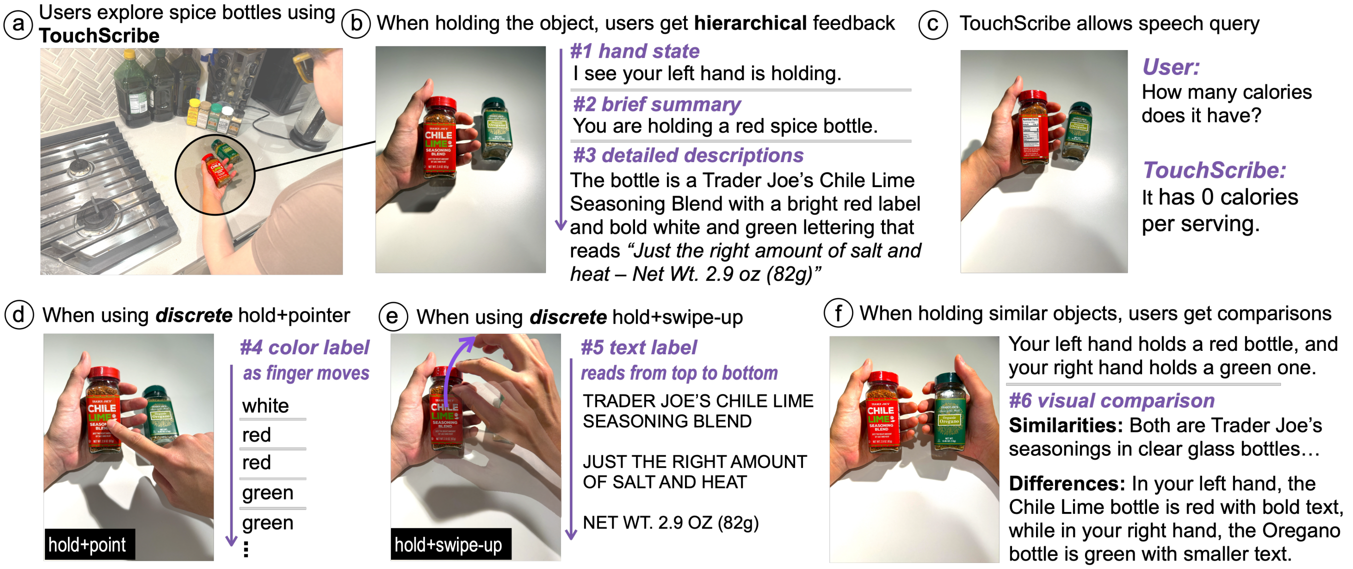
Wenqian Xu
University of Michigan
Ann Arbor, Michigan, USA
wtxu@umich.edu

Jovan Zheng Feng Yap
Computer Science and Engineering
University of Michigan
Ann Arbor, Michigan, USA
jovanyap@umich.edu

Tiange Luo
University of Michigan
Ann Arbor, Michigan, USA
tiangel@umich.edu

Venkatesh Potluri
School of Information
University of Michigan
Ann Arbor, Michigan, USA
potluriv@umich.edu

Anhong Guo
Computer Science and Engineering
University of Michigan
Ann Arbor, Michigan, USA
anhong@umich.edu



a Users explore spice bottles using **TouchScribe**
b When holding the object, users get **hierarchical** feedback
#1 hand state
I see your left hand is holding.
#2 brief summary
You are holding a red spice bottle.
#3 detailed descriptions
The bottle is a Trader Joe's Chile Lime Seasoning Blend with a bright red label and bold white and green lettering that reads *"Just the right amount of salt and heat – Net Wt. 2.9 oz (82g)"*
c TouchScribe allows speech query
User:
How many calories does it have?
TouchScribe:
It has 0 calories per serving.
d When using ***discrete*** hold+pointer
#4 color label
as finger moves
white
red
red
green
green
⋮
hold+point
e When using ***discrete*** hold+swipe-up
#5 text label
reads from top to bottom
TRADER JOE'S CHILE LIME SEASONING BLEND
JUST THE RIGHT AMOUNT OF SALT AND HEAT
NET WT. 2.9 OZ (82g)
hold+swipe-up
f When holding similar objects, users get comparisons
Your left hand holds a red bottle, and your right hand holds a green one.
#6 visual comparison
Similarities: Both are Trader Joe's seasonings in clear glass bottles…
Differences: In your left hand, the Chile Lime bottle is red with bold text, while in your right hand, the Oregano bottle is green with smaller text.

Figure 1: TouchScribe augments hand-object interactions with live visual descriptions. (a) BLV users use TouchScribe to explore objects placed on the shared kitchen counter. (b) While holding an object, BLV users receive hierarchical descriptions, starting with the hand state, followed by a brief summary, and then detailed descriptions. (c) They can also ask TouchScribe for visual information using speech. (d) Through finger gestures, users can access object details; for example, holding the object and pointing with the other hand reads its color; (e) holding and swiping up reads the text. (f) When holding two objects simultaneously, users receive a visual comparison highlighting their similarities and differences.

Abstract

People who are blind or have low vision regularly use their hands to interact with the physical world to gain access to objects' shape, size, weight, and texture. However, many rich visual features remain inaccessible through touch alone, making it difficult to distinguish similar objects, interpret visual affordances, and form a complete understanding of objects. In this work, we present TouchScribe, a system that augments hand-object interactions with automated live visual descriptions. We trained a custom egocentric hand interaction model to recognize both common gestures (e.g., grab to inspect, hold side-by-side to compare) and unique ones by blind people (e.g., point to explore color, or swipe to read available texts). Furthermore, TouchScribe provides real-time and adaptive feedback based on hand movement, from hand interaction states, to object labels, and to visual details. Our user study and technical evaluations demonstrate that TouchScribe can provide rich and useful descriptions to support object understanding. Finally, we discuss the implications of making live visual descriptions responsive to users' physical reach.

CHI '26, Barcelona, Spain
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791308

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI); Accessibility systems and tools**.

Keywords

Visual descriptions, blind, visually impaired, assistive technology, accessibility, hand-object interactions, gestures, LLM

ACM Reference Format:
Ruei-Che Chang, Rosiana Natalie, Wenqian Xu, Jovan Zheng Feng Yap, Tiange Luo, Venkatesh Potluri, and Anhong Guo. 2026. TouchScribe: Augmenting Non-Visual Hand-Object Interactions with Automated Live Visual Descriptions. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3772318.3791308

1 Introduction

People who are blind or have low vision (BLV) often rely on touch to explore and interact with the physical world, using their hands to perceive essential attributes of objects such as shape, size, weight, and texture [67, 99, 100]. However, many rich visual features remain inaccessible through touch alone, making it difficult to fully understand an object's appearance and functionality. For example, it is difficult to distinguish between grocery store products with similar shapes but different colors, patterns, or surface details without visual cues [60, 63]. Certain visual affordances, such as identically shaped seasoning bottles from different brands, similar tea bags with or without caffeine, or small printed ingredient labels, may be entirely imperceptible through touch.

Currently, BLV individuals can capture photos and engage in conversations with AI systems to obtain visual descriptions [7, 13, 18], or receive live narration to explore their surroundings [4, 34]. However, these AI systems often struggle to accurately identify the specific object of interest within an image (as noted in [62, 65]), generate long-form descriptions that contain unnecessary information, and their turn-taking interaction style can impede the quick retrieval of desired visual information. In contrast, hands provide a natural and intuitive interface for interacting with the physical world [61, 62, 64, 65]. Their movements are indicative of a person's intent [32] and objects of interest [33, 111], making them integral to how BLV individuals access and understand their environment through tactile exploration [41, 109]. Hence, in this paper, we investigate the question: *How can natural hand interactions be leveraged to support BLV people to access rich visual information about objects of interest?*

To address this question, we introduce TouchScribe, a system that provides live visual descriptions driven by hand-object interactions. TouchScribe supports a set of hand gestures inspired by prior research on discreet gestures preferred by BLV individuals [80, 81], such as pointing to an object held in the other hand to read its color [75, 89] or swiping across it to access available text (Figure 1d, e). TouchScribe also incorporates common gestures used with sight, such as touching or holding objects to signal interest (Figure 1b), and comparing two items side-by-side to explore visual similarities or differences (Figure 1f). Based on different hand interactions with objects, TouchScribe provides hierarchical feedback, including hand states for users to confirm that hand events are correctly identified, a brief object overview, and rich visual details (Figure 1b).

TouchScribe was prototyped using a neck mount with an attached smartphone (Figure 4). It detects fine-grained hand-object interactions, such as when both hands engage the same or different objects, or when an object is flipped, to deliver adaptive feedback that aligns with the user's evolving focus and intent. To support these hand-object interactions, we fine-tuned a custom egocentric hand gesture recognition model that interprets different hand gestures as information cursors to identify objects or information of interest. The underlying gesture recognition model is lightweight enough to run on live video feeds, and the wide camera field of view (FoV) provides broad coverage, though at the cost of accuracy due to inherent model limitations and distortion from the wide-angle lens. TouchScribe addresses these by integrating smoothing algorithms to mitigate intermittent recognition and extract keyframes. It also supports visual question answering (VQA), enabling users to freely query visual details when needed (Figure 1c).

We conducted a study with eight BLV participants to collect both qualitative and quantitative data, aiming to understand their experiences with using TouchScribe across various object-understanding tasks in our lab-controlled environment, and to evaluate the accuracy and latency of descriptions. Through qualitative analysis, we found that participants generally perceived TouchScribe interactions as intuitive (M=5.63 out of 7), with the provided descriptions being accurate (M=5.5) and comprehensive (M=6.5). Participants also felt a sense of control in the descriptions for object understanding (M=5.13). However, participants reported moderate cognitive effort (as measured by NASA-TLX) and a noticeable learning curve in hand positioning and gesture recognition with camera-enabled assistive technologies (ATs). Also, through our technical evaluation, we reported quantitative results to reflect TouchScribe's performance in our user study, including the accuracy of our custom hand posture recognition model in the live stream ($F_1 = 0.77$), the latency between detected hand movements and different types of descriptions (from 0.56s to 14s), and the accuracy of the descriptions (from 67.83% to 93.27%).

Through the study, we identified several gaps in the current TouchScribe prototype that limit its practical use. For example, while the wide camera FoV offered broader coverage, it also introduced inaccuracies due to image distortion. In addition, interpreting the intent behind diverse natural hand–object interactions remained challenging, and at times the system produced information overload during rapid hand gesture changes. Based on these findings, we discuss the implications to make TouchScribe generalizable for broader real-world situations in the future, such as customizations to different gesture preferences, integrating haptic-audio feedback for camera aiming, leveraging other gesture and object recognition techniques to improve accuracy, and making live visual descriptions responsive to users' further physical reach.

In summary, our work contributes:

(i) TouchScribe, a novel prototype system that generates live, rich object descriptions based on multiple hand-object interactions, moving beyond the single interaction and information types supported in earlier systems (Table 1).

Table 1: Overview of research and commercial apps for providing live visual descriptions leveraging different information cursors for real-world understanding.

Assistive System	Cursor Type	Information Type
Orcam [16], FingerReader [30], VizLens [43] and StateLens [44]	Finger tip	Text
Medeiros et al. [75] Stearns et al. [89]	Finger tip	Clothing color and texture
EyeRing [78]	Finger tip	Barcode, currency
SeeingAI [18]	Camera motion	Human, currency, barcode, object, color, lightness, and text
WorldScribe [34]	Camera motion	Object labels, general and detailed descriptions
TouchScribe (this work)	**Hands and fingers**	**Hand states, color, text, brief and detailed object descriptions, and object comparison**

(ii) A user study and technical evaluation demonstrating the intuitiveness of TouchScribe, usefulness of its descriptions, and its overall user experience.

(iii) Lessons learned from the development and evaluation of TouchScribe, and design implications for employing egocentric camera-enabled, real-time assistive technologies in the real world.

2 Related Work

Our work builds upon and connects three key research domains. First, prior research on hand-based interactions has shown that the expressive and intentional nature of hands provides a compelling alternative to device-based input (e.g., controllers) though user preferences vary depending on different contexts, which informed our selection of gestures in TouchScribe. Second, studies on the use and limitations of hand interactions in current ATs for accessing real-world information revealed opportunities for TouchScribe to incorporate more expressive hand–object interactions and deliver richer visual information. Third, advances in vision–language models (VLMs) have demonstrated their potential to enhance access to visual content without human assistance; however, they remain limited in usability and in providing live object descriptions driven by hand–object interactions. This motivated our approach of using hands as information cursors to proactively deliver essential visual information beyond the repetitive speech prompts of current AI-enabled ATs. Below, we discuss insights from these domains that shaped the design of TouchScribe.

2.1 Hand Interactions as Intent Cues

Hands provide a natural and intuitive interface for interacting with the physical world, effectively conveying users' intentions [32], actions [72], and objects or areas of interest [33, 111]. Hand gestures are highly expressive and support a wide range of tasks for the general population, including animation creation and authoring [28, 70, 85, 110], mode switching [91], typing [56, 105], and object manipulation [51, 58, 66, 76, 82, 97, 106]. Beyond visual interactions, hands also play a crucial role in nonvisual exploration. For BLV individuals, tactile exploration strategies vary widely and include bimanual, unimanual, and alternating approaches [98, 109], which demonstrated the adaptability of hand use strategies to different information needs. However, when considering the social acceptability of hand interactions, on-body gestures performed within the hands, such as tapping or swiping a finger across one's opposite palm, are generally preferred. Unlike bodily gestures (e.g., making an 'OK' sign, waving) [37], these gestures are more discreet, socially acceptable, and feel natural in everyday contexts, such as quickly checking for new messages while commuting [80, 81]. Drawing from these works, in TouchScribe, we also considered unique and usable hand interactions for accessing information.

2.2 Current Use of Hand Interactions for Assistive Technologies

Hand-based interactions have been explored in both commercial ATs and prior research. For instance, BLV individuals commonly access digital information through touch gestures on smartphones. Swipe gestures, for example, enable screen navigation, such as swiping left or right for word-by-word reading, or using a two-finger swipe up in screen readers like TalkBack [10] or VoiceOver [19] to read from the top of the screen. While these methods are effective in digital contexts, comparable approaches for accessing information of physical objects remain limited, often requiring photo capture followed by a question–answering process. Though tactile exploration can support object understanding [41, 109], many rich visual features, such as labeled texts, colors, or intricate patterns, remain inaccessible through touch alone.

To bridge this gap, prior research has explored using the hands and fingers as information cursors to access visual information in real-time [45]. For instance, prior systems, such as VizLens [43], StateLens [44], and FetchAid [42], support interactions with appliance control panels by allowing users to point to interface elements that are subsequently read aloud. Finger-mounted camera systems have been explored as a means of supporting BLV users in accessing visual details, including text [30, 40, 86, 88], currency and barcodes [78], and clothing color and texture, while maintaining tactile feedback for hands-on exploration [75, 89]. Building on this direction, Lee et al. [61, 62, 64, 65] proposed custom models that leverage hand position to localize objects of interest for more effective intent disambiguation and camera alignment.

Despite these advances, existing systems (Table 1) often rely on a limited set of gestures and hand-held devices for photo capturing, and provide only single or limited forms of visual feedback (e.g., text, color). In contrast, enabled by an integrated hand recognition and description generation pipeline, TouchScribe offers a fluid, hands-free, and integrated experience by delivering live, rich object descriptions driven by hand-object interactions. For instance, BLV users can hold or touch an object with one hand to obtain rich visual details, point the object with another to read colors, or perform a swipe-up gesture to access its available texts. Such natural and expressive information access was lacking in prior systems.

2.3 Visual Descriptions with VLMs

Beyond relying on remote sighted assistance [1, 2] or crowdsourcing [29, 92], where human agents may not always be available, recent advancements in VLMs have enabled applications that allow BLV individuals to easily submit image-description queries to AI-powered VQA systems [3, 7, 13, 59, 102] or receive real-time visual descriptions from live video AI systems [4, 34]. These technologies promote the independence and autonomy of BLV individuals without requiring sighted assistance. We discuss them below.

2.3.1 *Image Capture and Visual Question Answering.*

Current AI-powered visual description systems require users to capture photos and engage in dialogue with AI assistants to obtain specific visual details [3, 7, 13, 59, 102]. This process of photo capturing and turn-taking can be laborious and time-consuming. For example, taking pictures demands precise camera alignment to ensure the object of interest is within the frame [22, 29, 53, 54, 93], often involving repeated trial and error. Although cameras with a wider FoV may help mitigate this issue [49], users must still explicitly specify their needs and interact with the AI to obtain desired details. This turn-taking VQA process is further challenged by the dynamic nature of the real world, where generated descriptions can quickly become outdated as the environment changes.

2.3.2 *Live Video Feed and Generative Descriptions.*

Building on photo-taking, ChatGPT's Advanced Voice with Video [4] enables a conversational approach to retrieving visual information through a live video feed. However, instead of proactively delivering essential details, it depends on continuous speech prompts from the user [35], which may introduce turn-taking delays, increase effort, and raise concerns about privacy and social acceptability. To overcome this lack of proactivity, WorldScribe [34] provides live visual descriptions that dynamically adapt to camera motion and the captured visual content. For example, WorldScribe [34] enhanced users' environmental awareness by providing brief object labels as the camera panned across the surroundings, and offered richer visual details when the camera focused on a specific scene. In contrast to environmental understanding, our work explores using *hand gestures* as information cursors to proactively describe objects based on how the user is interacting with them, enabling more responsive, intuitive, and fine-grained *object understanding* in real time.

3 TouchScribe

TouchScribe is a system that provides live, rich object descriptions based on the user's hand interactions with physical objects. It detects three types of hand gestures and identifies hand activities in each frame (*Hand Gesture Recognition Layer in Section 3.4*). Then, TouchScribe extracts keyframes from live video stream based on these hand activities (*Keyframe Extraction Layer* in Section 3.5) to generate multiple forms of feedback, including hand states, object color, available texts, and brief and detailed object descriptions, and object comparisons (*Description Generation Layer* in Section 3.6). We describe our design goals and implementation details below.

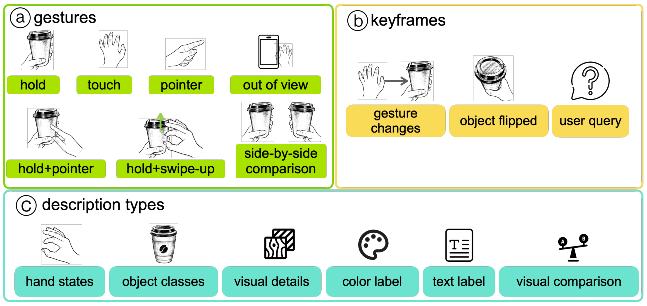


Figure 2: Overview of the variety of gestures, timings to extract keyframes, and description types supported by TouchScribe.

3.1 Design Goals

TouchScribe is designed based on three primary goals inspired by prior work:

G1 - Supporting common and usable gestures. Hand interactions serve as valuable intent cues for disambiguating objects of interest [61, 62, 64, 65] and indicating the locations of relevant information (e.g., text [42–44] or color [75, 89]). However, because hand–object interactions vary across individuals and contexts [109], as an initial step, TouchScribe should demonstrate a set of common and usable gesture types for BLV individuals.

G2 - Supporting proactive and real-time feedback. Given the current strengths and limitations of photo-capturing and VQA-based approaches [35, 101], which provide access to specific information but introduce delays and turn-taking overhead, TouchScribe should primarily emphasize proactive feedback while still supporting VQA when needed. Moreover, TouchScribe's descriptions should be closely synchronized with hand interactions, minimizing latency between touch and audio output to enhance the overall user experience.

G3 - Conveying system-perceived states of hand-object interactions. Given that camera aiming has long posed challenges for BLV users [22, 29, 53, 93], it is essential to clearly communicate whether the system has detected users' hands and what it has recognized, enabling users to take appropriate follow-up actions.

3.2 Gestures to Access Visual Information

To fulfill **G1**, TouchScribe supports six gestures (Figure 2a), categorized along two dimensions: ***(i) familiarity***, gestures that are *common* versus those *unique* to BLV users, and ***(ii) gesture nature***, gestures that are *continuous* versus *discrete*. These gestures are informed by prior research on commonly used assistive technologies or discreet on-body gestures [80, 81]. Each gesture maps to a distinct prompt for VLMs to generate corresponding descriptions (Details are in Section 3.6).

(1) **Hold an object with a single hand *(common & continuous)*** – Holding an object of interest is a common practice for examining its visual or tactile details, such as reading nutritional information on a bottle or exploring its shape.
(2) **Touch an object with a single hand *(common & continuous)*** – Touching an object with a few fingers is common for indicating an object of interest in sighted interaction [111], and is also widely used in tactile exploration by BLV people [109].

(3) **Hold or touch explore an object with both hands *(unique & continuous)*** – Using both hands to explore objects through touch is a common tactile exploration pattern among BLV individuals, particularly when interacting with flat or textured surfaces such as tactile graphics [41, 109].
(4) **Hold or touch objects side-by-side with both hands *(common & discrete)*** – When comparing similar items, such as ingredient labels on two bottles or subtle shape differences between boxes, people often place or hold them side by side to facilitate comparison.
(5) **Hold an object in one hand and point with another hand to reveal visual details *(unique & discrete)*** – Pointing gestures are common for BLV people to access specific information, such as color [89], text [30, 40, 43, 44, 86–88], or texture [89, 104].
(6) **Hold an object in one hand and two-finger swipe up with another hand to read texts. *(unique & discrete)*** Two-finger swipe-up gestures are commonly used in screen readers such as iOS VoiceOver [19] and Android TalkBack [10] to read on-screen text from top to bottom. Because the exact locations of text are often unknown to BLV people, we adapt this gesture to enable access to available text on an object.

3.3 Implementation Details

To enable TouchScribe to provide real-time feedback (**G2**), we trade off different factors to maximize the computing speed while maintaining decent accuracy (See Section 6), such as the choices of the models, or the frame size. TouchScribe servers include a local server running on a MacBook M4 Max and a remote server with two embedded Nvidia GeForce RTX 4090 GPUs. TouchScribe uses a neck mount with an attached iPhone 13 Pro (Figure 4). The smartphone offers more APIs than emerging smart glasses at the time of development, and greater flexibility in selecting frame resolution for real-time use and camera FoV for coverage. The TouchScribe iOS app uses the wide lens, the 13 mm-equivalent rear camera with an approximately 120° FoV, whereas the standard wide lens (26 mm-equivalent) offers a 77° FoV. It streamed the video frames (width: 720, height: 960, configured to retain approximately 70% image quality) to the local server through a Socket connection. Google MediaPipe [9], the hand gesture recognition model and finger motion classification model, runs on the local server and achieves around 6 frames per second (FPS). Other models that require higher computational resources run on the remote server, including Hands23 [36] for detecting hand–object contacts (F_1-score=79.1), SigLIP [107] for generating image embeddings, and Moondream [15] for producing brief object descriptions.

3.4 Hand Gesture Recognition Layer

In this layer, TouchScribe aims to recognize the aforementioned hand gestures in a lightweight manner to support real-time performance alongside other models for live visual descriptions (**G2**). To achieve this, we fine-tune a hand gesture classification model and a finger motion classification model, which identify gestures and finger movements based on hand landmarks detected using Google MediaPipe [9]. We describe these models in detail below.

3.4.1 *Hand gesture classification model.*

We fine-tuned a publicly available keypoint classification model [12] to adapt its model structure for our supported gesture set. The model takes a 2D keypoint vector as input and outputs a hand gesture class. Specifically, the input vector has a dimensionality of 42 corresponding to the (x, y) coordinates of 21 hand keypoints extracted from Google MediaPipe [9]. The output includes three gesture categories, *touch*, *hold*, and *point*, for each hand. Additionally, a gesture is labeled as *out of view* when no hand keypoints are detected by Google MediaPipe [9]. This results in a total of four classes for each hand.

3.4.2 *Finger motion classification model.*

To support the two-finger swipe-up gesture described in Section 3.2, we fine-tuned a finger motion classification model from the same publicly available repository [12]. The model takes as input a time-series history of a fingertip's 2D coordinates (x, y), sampled every 16 frames and resulting in a flattened input vector of size 32. The output includes three finger motion gesture categories, including *static*, *up*, and *down*. This model is executed only when one hand is in the *hold* state and the other is in the *touch* state, with both the index finger and thumb of the *touch* hand located within the bounding box of the *hold* hand (Figure 3e).

3.5 Keyframe and Object Extraction Layer

In this layer, TouchScribe identifies keyframes when users perform new gestures or flip an object, signaling the need for updated descriptions (**G3**). A key challenge arises from the intermittent predictions produced by the gesture recognition models due to real-time performance requirements (**G2**), which may reduce accuracy and lead to false positives. To mitigate this issue, TouchScribe applies a temporal smoothing function that analyzes consecutive frames to infer a stable gesture state for each hand.

First, TouchScribe verifies whether the past x gestures of either hand consist of a single gesture class repeated at least t times (e.g., *hold*, *touch*, *point*, or *out of view*). If this condition is satisfied, that gesture is assigned as the stable gesture state. Otherwise, TouchScribe checks whether the previous stable gesture state appears within the last n frames and retains it if so. If neither condition holds, the most frequent gesture in the last n frames is selected as the current gesture. Based on our apparatus and empirical tests, we set $x = 12$, $n = 6$, and $t = 4$. Whenever the stable gesture state transitions to either *hold* or *touch*, the corresponding frame is marked as a keyframe and sent to the Hands23 [36] model to identify hand–object contact details, supplementing prompt data for the description generation pipeline (Figure 3b).

In addition, when the stable gesture of either hand remains as *hold* or *touch* across keyframes, TouchScribe analyzes whether the object is unchanged by periodically cropping the object image and computing the cosine similarity between the current image embedding and those from the previous s samples (Figure 3f). If the similarity scores with all s prior samples fall below a threshold u, the frame is marked as a keyframe, which indicates a potential change or flip of the object. Based on our apparatus and empirical tests, we set $s = 4$ and $u = 0.85$.

The extracted keyframes and objects are passed to VLMs to generate hierarchical feedback and descriptions. The structure of these prompts and outputs is detailed in the following section.

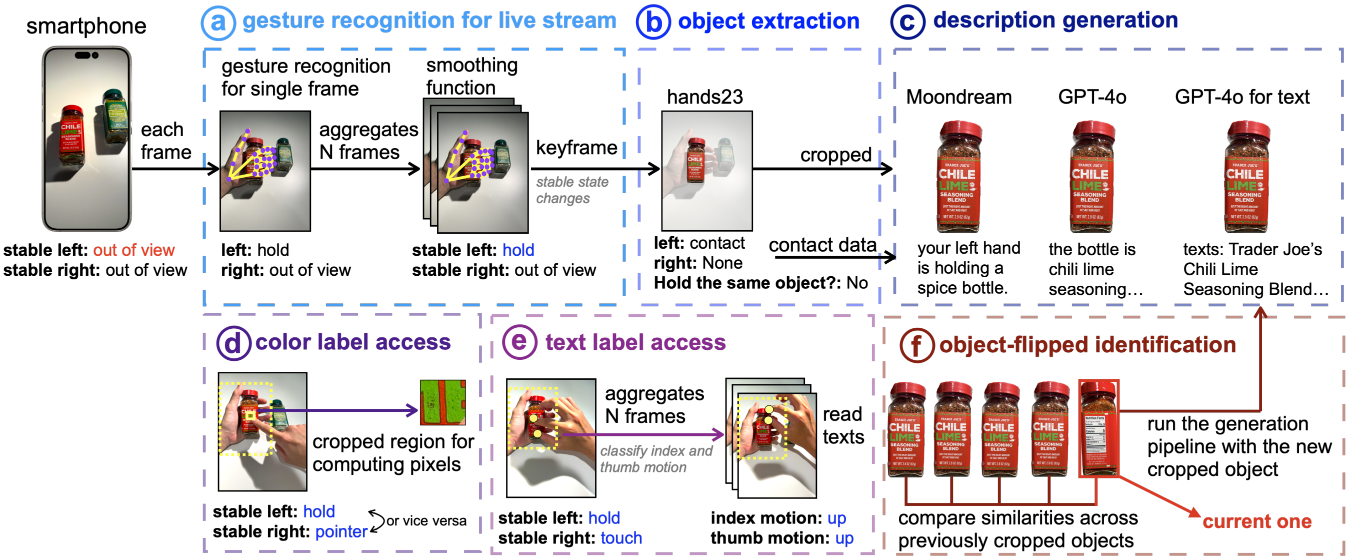

smartphone
each frame
stable left: out of view
stable right: out of view
a gesture recognition for live stream
gesture recognition for single frame
aggregates N frames
smoothing function
keyframe
stable state changes
left: hold
right: out of view
stable left: hold
stable right: out of view
b object extraction
hands23
cropped
left: contact
right: None
Hold the same object?: No
contact data
c description generation
Moondream
GPT-4o
GPT-4o for text
your left hand is holding a spice bottle.
the bottle is chili lime seasoning…
texts: Trader Joe's Chili Lime Seasoning Blend…
d color label access
cropped region for computing pixels
stable left: hold
stable right: pointer
or vice versa
e text label access
aggregates N frames
classify index and thumb motion
read texts
stable left: hold
stable right: touch
index motion: up
thumb motion: up
f object-flipped identification
run the generation pipeline with the new cropped object
compare similarities across previously cropped objects
current one

Figure 3: TouchScribe System Diagram. (a) TouchScribe performs gesture recognition on live video streams. For each camera frame, hand landmarks are extracted with Google MediaPipe [9] and classified into predefined gesture categories. A temporal smoothing module then aggregates multiple frames to produce stable keyframes and gesture states. (b) For each keyframe, Hands23 [36] infers object contact. The contact data, together with a cropped image of the object, is passed to VLMs for further processing. (c) VLMs, including Moondream [15] and GPT-4o [11], are executed in parallel to generate rich object descriptions. (d) When one stable state is *hold* and the other is *point*, TouchScribe reads the color of the small region the finger is pointing to. (e) When one stable state is *hold* and the other is *touch*, TouchScribe tracks finger motion and reads the text once both fingers move up. (f) TouchScribe also maintains a history of cropped objects and identifies flipped instances by comparing image similarity, and re-runs the generation pipeline on the updated crop.

3.6 Description Generation Layer

In this layer, TouchScribe generates descriptions with adaptive levels of detail based on the user's hand-object interactions. To achieve this, TouchScribeintegrates outputs from multiple components, including gesture states from the hand gesture recognition models, hand–object contact information inferred by the Hands23 [36] model, and the extracted keyframes and objects. TouchScribe dynamically incorporates these details to construct descriptions or prompts for VLMs to generate rich object details. We detail each type of description and its generation process below.

Hand-State Feedback. Hand-state feedback helps users assess whether the hands are correctly captured and identified (**G3**). Whenever the user's hands are detected within the camera view, TouchScribe generates feedback such as "I see your {which_hand} hand" to help users confirm the presence of their hands in the frame, where "{which_hand}" is dynamically assigned as *left*, *right* or *both*. Then, TouchScribe describes the perceived stable gesture states, for example: "Your {which_hand} hand is/are {gesture}ing" or "You flipped or changed the object.", where "{gesture}" is dynamically assigned based on the recognized stable gesture state, including *hold*, *touch* and *point*. The two feedback are combined when they are temporally close to reduce repetition, such as "I see your right hand is pointing."

Brief Object Descriptions. The brief description helps users quickly assess what the object is, whether it is of interest, and whether they want to learn more. Given a keyframe, TouchScribe first applies the Hands23 model [36] to obtain hand–object contact information, including which hand (or both) is in contact and a cropped image of the object (Figure 3b). When contact is detected, TouchScribe generates prompts such as "What is my {which_hand} hand touching?" with a cropped object image. This prompt is then passed to Moondream [15], a lightweight VLM that produces concise descriptions with low latency. Example outputs include "Your right hand is touching a bottle of seasoning." and "Both your hands are touching a laptop."

Detailed Object Descriptions. Detailed descriptions enable users to access fine-grained visual information about objects. Using the same cropped object image and hand–object contact data provided to Moondream [15], TouchScribe also supplies these inputs to GPT-4o [11] with a different prompt: "Can you describe the object I am {gesture}ing with my {which_hand} hand in detail?" This produces descriptions such as "You are holding a white mug decorated with colorful illustrations..." Although Moondream [15] and GPT-4o [11] perform inference in parallel, Moondream generates an initial high-level description first, followed by GPT-4o's more detailed output due to differences in latency.

Available Object Texts. TouchScribe reads aloud the available text on the object (e.g., expiration date, nutrition facts) once the user performs the *hold+swipe-up* gesture. Using the same cropped object image, TouchScribe submits a different prompt to GPT-4o [11]: "I am holding the object with my {which_hand} hand. Please describe the text line by line. If there is no text, can you just return 'no text on the {object name} your {which_hand} hand is {gesture}ing." We employ GPT-4o [11] for its acceptable latency and accuracy of text recognition on low-resolution images compared to other text recognition models. This approach enables top-to-bottom reading of text on object surfaces, analogous to screen readers such as iOS VoiceOver [19] and Android TalkBack [10]. If users trigger the gesture before texts are generated, TouchScribe responds: "Still processing the text, please try again later."

Comparative Descriptions. This feedback aims to support comparisons between objects with similar tactile features (e.g., shape and size), enabling users to better understand their visual similarities and differences. When both hands *hold* or *touch* different objects, TouchScribe crops the corresponding object images using the Hands23 model [36] and prompts GPT-4o [11]: "Can you describe the object I am holding with my left hand and the one with my right hand? What are the differences or similarities between them?" This yields outputs, such as "Your left hand holds a red bottle, and your right hand holds a green one. Similarities: Both are Trader Joe's... Differences: color and texts are different ..." Also, TouchScribe detects when both hands *hold* or *touch* different parts of the same object, supporting users in understanding the object's spatial layout and visual characteristics (e.g., surface graphics and text), building on prior work [41, 109]. In this case, TouchScribe prompts GPT-4o [11] with full image and instructions: "Can you describe the spatial and visual relationship between the points I am touching, and highlight any visual similarities or differences between them?" Example outputs include "Your hands touch adjacent areas around the bottle, with the left spanning the text... and the right spanning the graphics..."

Color Labels. TouchScribe reports an object's color when users *hold* it with one hand and *point* to it with the other. Then, TouchScribe analyzes a small image region near the index fingertip. Based on the fingertip coordinates and hand side (`left` or `right`), the system slightly offsets the cropped region (left/up for the right hand and right/up for the left hand) to exclude the finger itself. It then computes the region's average RGB value and maps it to the nearest named color using the *webcolors* library [20].

User Query. Lastly, in line with **G2**, TouchScribe enables users to invoke a question-answering function via the voice command "Hey <wake word>" and pose queries. TouchScribe then submits the current video frame along with the user's question to GPT-4o [11] and reads the generated response, similar to existing AI-enabled assistive VQA services such as BeMyAI [13] and SeeingAI [18].

3.7 Handling Responsiveness of Descriptions to Hand Interactions and Speech Query

TouchScribe prioritizes descriptions and interrupts based on different hand–object gestures. For example, invoking the VQA function interrupts any ongoing narration to address the query, after which hand gestures are ignored until the answer is fully delivered. In contrast, discrete gestures for specific visual information, such as *hold+point* for color labels or *hold+swipe-up* for object text, can also interrupt ongoing descriptions.

4 Evaluation Methods

We conducted a user study to qualitatively understand **How do BLV participants experience and perceive TouchScribe?** We then used the captured videos and interaction data from the study to conduct a technical evaluation for quantitative insights into **What is the accuracy and latency of TouchScribe's descriptions, in response to users' hand-object interactions?** We detail our methods and results below.

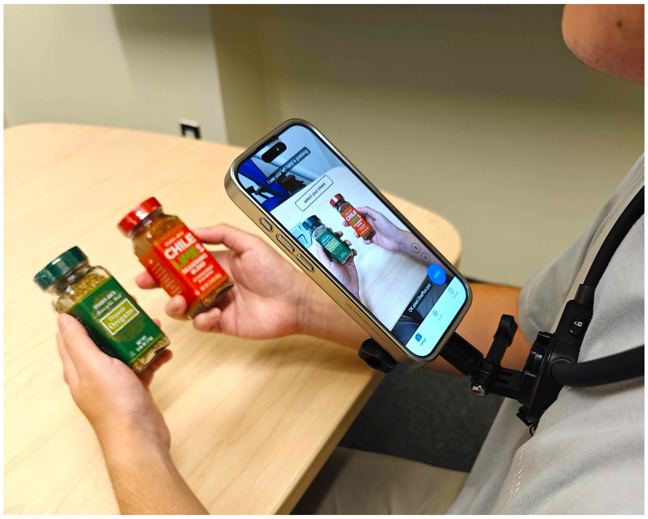


Figure 4: The TouchScribe prototype setup included an adjustable neck mount with an attached smartphone. During the study, researchers adjusted the mount for each participant to ensure the camera was properly aimed at the table.

4.1 Participants

We recruited eight BLV participants (3 Male and 5 Female) using email lists for local accessibility organizations, prior contacts, and snowball sampling. Participants aged from 18 to 72 (Avg. 45.5) and described their visual impairment as blind (N=6) or having low vision (N=2). Most participants had prior experiences using remote sighted assistance and AI-enabled services, such as Orcam [16], BeMyEyes [2], BeMyAI [13], Aira [1], or SeeingAI [18] in their daily lives (Table 3).

4.2 Procedure, Apparatus and Tasks

The study consisted of two sessions: (i) a *practice session*, designed to familiarize participants with TouchScribe, and (ii) a *task session*, during which participants completed a series of object understanding and selection tasks. Throughout the study, participants remained seated and interacted with the system using a neck-mounted smartphone (Figure 4).

(i) Practice session. Participants were introduced to the hand gestures supported by TouchScribe, the corresponding feedback and descriptions provided by the system, and the procedure for invoking the VQA function.

(ii) Task session. During the task session, participants completed four object understanding and selection tasks with increasing levels of complexity, determined by the number of objects and the specificity of required information [35]. We describe each task below.

(1) **Understanding an object**: Participants were given a cup featuring text and graphics on its surface (Table 2). The cup was placed on a table, and participants were instructed to use TouchScribe to obtain descriptions to understand its visual features. The task concluded when participants felt they had sufficiently understood the cup's visual characteristics and reported their observations to the experimenter.

Table 2: Setup and instructions for each scenario. These scenarios differed based on factors such as Visual Complexity in object understanding tasks marked as *Low*, and *High* in purple, and Information Specificity in blue (e.g., *Specific* vs. *General*).

Image	Scenario	Setup	Instruction to User	Dimensions
	Understanding an object	Participants were given a cup with colorful graphics and texts.	You got a gift from your friend who just traveled back from a tourist spot. Can you use TouchScribe to understand this object? In terms of color, texts, and graphics.	**General** **Low**
	Understanding and distinguishing two different spice bottles	Participants were given two spice bottles from Trader Joe's, including one chili lime seasoning with a red label and lid, and another oregano with a green label and lid.	In the grocery store, you have two spice bottles with different labels, colors, and texts. Can you use TouchScribe to tell the differences and the similarities between them?	**General** **High**
	Understanding and categorizing four spray bottles	Participants were given two identical (from the brand *Everyone*, ruby grapefruit), and the other two were from the same brand (*Whole Foods 365*) but had different scents (cucumber aloe and lavender).	You just got the four spray bottles from a shared storage in your home. Can you use TouchScribe to categorize them based on their brands and scents?	**Specific** **Low**
	Finding products with specific information	Participants were given three carton of juices, including two apple juices (100 & 35 calories) and one lemonade (100 calories), and three chocolate bars (55, 65, 70% of cocoa).	You want to find some snacks in a shared pantry, specifically, the chocolate bars with the most cocoa and the apple juice with the fewest calories for your health. Can you use TouchScribe to help you find them?	**Specific** **High**

(2) **Distinguishing two similar objects**: Participants were provided with two seasoning bottles of identical shape but differing in labels, colors, and text. They were asked to identify both similarities and differences between the bottles. The task concluded when participants felt they had sufficiently understood these attributes and reported their observations to the experimenter.

(3) **Sorting four similar objects**: Participants were provided with four bottles of similar shape and size: two identical bottles of the *Everyone* brand (grapefruit scent) and two bottles from the *Whole Foods 365* brand with different scents (cucumber aloe and lavender). They were asked to categorize the bottles by brand and scent. The task concluded when participants indicated they had completed the categorization.

(4) **Selecting objects based on specified needs**: Participants took part in a shared pantry scenario in which they were asked to locate items based on specific nutritional information. The setup included six products: three chocolate bars with varying cocoa content (55%, 65%, and 70%) and three beverages, two apple juices with 100 and 180 calories, and one lemonade. Participants were instructed to identify the chocolate bar with the highest cocoa content and the apple juice with the fewest calories. The task concluded when participants indicated they had finished.

For each task, objects were randomly placed on the table in front of participants rather than deliberately staged. This allowed the objects to be encountered naturally without excessive search time, as object finding was not the focus of our study. To support the collection of qualitative insights, participants were encouraged to think aloud and take their time exploring TouchScribe while completing the tasks. After completing all tasks, participants responded to a set of Likert-scale questions (Figure 5), completed the NASA-TLX form to assess cognitive load (Figure 6), and shared their overall experiences.

The entire study lasted about one hour. Participants were compensated for their transportation costs and an additional $25 for their participation. This study was approved by the Institutional Review Board (IRB) at our institution.

4.3 Data Collection and Analysis

For the user evaluation, we collected participants' responses to a set of Likert-scale questions across multiple dimensions, including perceived effectiveness, intuitiveness, usefulness, perceived accuracy and coverage of descriptions, and sense of agency when using TouchScribe (Figure 5). Participants also completed the NASA-TLX questionnaire [47] to assess cognitive workload. Additional insights were obtained through open-ended questions in a semi-structured interview, and the entire session was video recorded. Two researchers transcribed the interviews and analyzed the qualitative data using affinity diagramming.

In addition, interactions with TouchScribe were logged for technical evaluation, including recognized gestures, generated descriptions, and referenced frames (Section 6). To analyze these data, we conducted a round-table discussion and annotation session with four members of the research team. The researchers collaboratively reviewed the images and their corresponding descriptions, dividing the workload. Ambiguities or questions raised by any team member were resolved through group discussion.

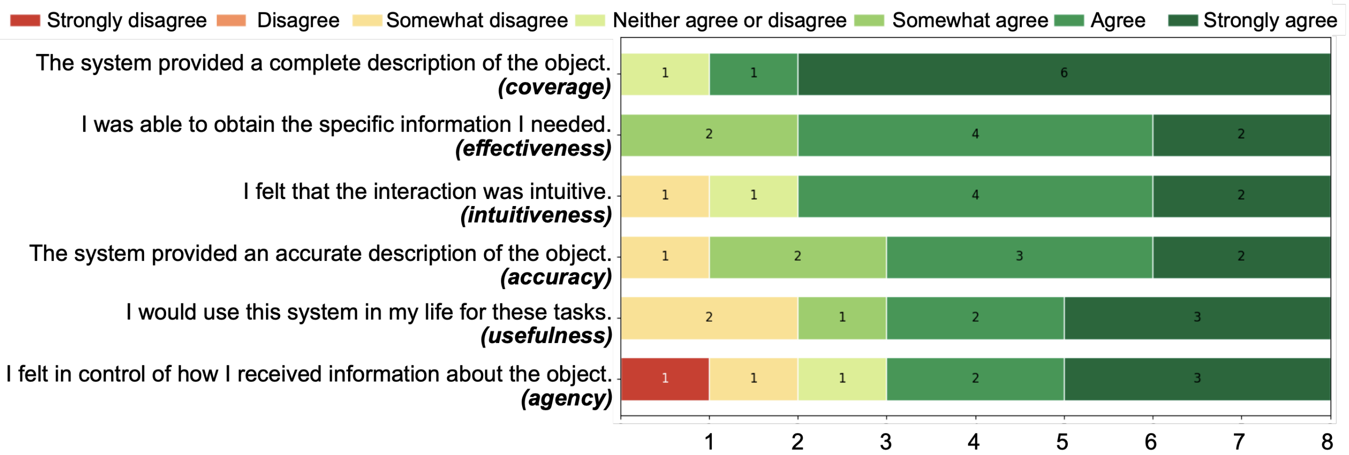


Figure 5: Likert scale questions and aggregated responses of eight participants in our user study. This includes questions about coverage (M=6.5, SD=1.07), effectiveness (M=6, SD=0.76), intuitiveness of gestures (M=5.63, SD1.41), accuracy of descriptions (M=5.5, SD=1.6), usefulness (M=5.5, SD=1.69), and agency of using TouchScribe (M=5.13, SD=2.23).

4.4 Limitation

Our study was conducted in a controlled lab environment, with participants seated throughout the sessions due to the study's extended duration. We acknowledge that this setting may not fully reflect real-world conditions, where users may interact with cluttered environments or objects that exceed typical hand-grasp ranges. Although the neck-mounted smartphone was designed to approximate an egocentric perspective, it may be impractical for everyday use because of potential social acceptability concerns. Such issues could be mitigated through alternative form factors, such as smart glasses or more discreet wearable setups (e.g., a yarn lanyard). Additionally, lighting conditions and camera angles were adjusted for each participant to accommodate the limitations of the current hand landmark detection model. Despite these constraints, our primary goal was to demonstrate the feasibility of delivering live descriptions driven by hand-object interactions. We discuss these limitations and potential solutions in Section 7.

5 User Evaluation Results

In general, participants were able to use TouchScribe to complete a majority of the tasks. They commended the accuracy and coverage of the information provided, as well as the intuitive way to access specific details, especially in comparison to the tools they currently use. However, participants also identified several limitations, including latency in retrieving specific information due to the hierarchical feedback design, interruptions triggered by unintentional hand movements, and a learning curve associated with the new interaction techniques. We elaborate on these findings below.

5.1 Overall Task Completion

Overall, participants successfully completed the majority of tasks (27 out of 32 tasks), typically within 5-10 minutes, and reported high perceived effectiveness in using TouchScribe to obtain specific information (M=6.0, SD=0.76).

Specifically, for *Task 1 – Understanding an object*, participants achieved an 87.5% completion rate by correctly identifying the text and colors on the cup. Most used gestures, such as *hold* the cup and flip it around to access surface details, and use and *hold+point* to access its color. One exception was P1, who misidentified the interior color as black due to shading while pointing inside the cup.

For *Task 2 – Distinguishing two similar objects*, all participants successfully identified differences in brand names, spice labels, and bottle colors. Common strategies we observed included holding both bottles side-by-side for comparative descriptions or examining each bottle individually at a time to verify visual details.

Similarly, in *Task 3 – Sorting four similar objects*, participants reached a 75% completion rate. Most participants distinguished the bottles using both color labels and visual descriptions, but some struggled with reading text on curved surfaces, leading to hallucinated or incomplete descriptions. This caused confusion for P2 and P6, who did not complete the task.

In *Task 4 – Selecting objects with specified needs*, the completion rate was 75%. All participants successfully identified the chocolate bar with the highest cocoa content, but some encountered difficulties with the juice selection. For example, P1 was unable to locate the side with the calorie label and gave up, while P3 misremembered the calorie values despite receiving accurate descriptions.

5.2 Perceived Accuracy, Completeness, and Latency of Descriptions

Participants found that TouchScribe provided accurate and comprehensive descriptions; however, the density and prioritization of the information occasionally hindered efficient access.

Overall, participants perceived descriptions to be accurate (M=5.5, SD=1.6) and complete (M=6.5, SD=1.07), such as *"I can get descriptions of bottle, texts on it, and colors too. Without a person or an app like Be My Eyes or Aira, you usually just get one of them and miss the full picture."* (P3) or *"It's detailed, descriptive, and reads ingredients verbatim per se"* (P2). P1 also found the coverage of TouchScribe's descriptions informative than his current apps: *"If I use Seeing AI, I just held it there a minute until it starts reading* *words. And as soon as I recognized a keyword, I knew what it was. Whereas with [TouchScribe], it is more. It doesn't just read the 1st word that it comes to, but also recognizes the object like a box of cereal. It's Cheerios, whereas Seeing AI is just gonna start reading randomly, heart health, 100 calories, and great with milk, and then it might say, Cheerios. [TouchScribe] is recognizing the object, instead of just saying words."*

However, despite the comprehensive coverage of information, participants expressed mixed perceptions regarding the density and prioritization of the spoken content. For example, participants noted that hand-state feedback would be more appropriate as *"a tutorial at the beginning to understand what it sees (P6),"* rather than being presented regularly, which they found somewhat distracting. Also, P4 felt the transitions between descriptions were smooth, but suggested adding a brief pause in between for easier comprehension: *"It was telling me more than what I needed to know at that moment. Maybe consider adding a second or two."*

Furthermore, the hierarchical feedback design, progressing from brief to more detailed object descriptions, presented both advantages and drawbacks. On the one hand, it could slow access to specific details, such as retrieving nutritional information in Task 4, where a direct VQA might be more efficient. On the other hand, it helped contextualize information and maintain coherence across descriptions. For example, P3 found the hierarchical structure helpful for distinguishing between similar objects, noting in Task 3: *"It said that all of them were hand sanitizers and then went down into more specific information, like this is orange blossom, and this is cucumber. It drills down to the more specific information, and it would be easy to tell which is which."*

5.3 Perceived Agency, Gesture Intuitiveness, and Hand Constraints

Participants generally found the gestures intuitive and felt in control when accessing information, though they also noted usability challenges related to hand movements.

Participants rated the gestures as intuitive (M=5.63, SD=1.41) and reported a sense of control when using TouchScribe (M=5.13, SD=2.23). P2, who regularly used OrCam [16] for text reading, noted that hand-based interaction provided greater control: *"The way you move your hand tells the system everything it needs to describe the object. For glasses, you have to chin down, use your nose, and go down towards the text to get everything in the block. This (TouchScribe) did me a replacement by just holding the object."*

Similarly, P4 appreciated the immediacy of TouchScribe compared to applications used in daily life, such as Be My AI [13]: *"I just want immediate responses, because Be My AI will take a picture and tell you some basic things. And you need to go to chat for more information. [TouchScribe] is more immediate. You don't have to go through a chat to do it. That's just right there at the fingertips. Immediate. This would be a good app for people who do not have the patience to mess with chat."*

Participants also found TouchScribe's feedback on text detection and object flipping helpful for identifying items whose visual information is distributed across multiple surfaces and not fully visible from a single viewpoint. This feature reminded participants of grocery shopping experiences in which they needed to locate product barcodes to access digital information using existing assistive applications (e.g., Seeing AI [18]), suggesting that TouchScribe could provide practical benefits in such contexts. As noted by P3, *"That's difficult if you don't know where the barcode is located. You need to turn the item in all kinds of ways to get the system to recognize that barcode. With this (TouchScribe), you don't need to wait for locating the barcode. It just told what this is, and if there is text. So I knew to turn it to the other side."*

During the study, TouchScribe occasionally misinterpreted idle hand movements or noise in posture detection as intentional input, resulting in false positives and unintended interruptions (see Section 6.1). Participants noticed these disruptions but generally viewed the system's sensitivity as a trade-off. As P7 said: *"It restarted whenever I was even just moving a little bit... but checks and balances...because previous descriptions might go on for too long if it didn't restart."* To adapt, some participants intentionally moved their hands out of the camera's view to pause the system and then brought them back to reset the hierarchical feedback. We further discuss these limitations and propose future directions for improving gesture recognition and intent disambiguation in real-time settings in Section 7.2.

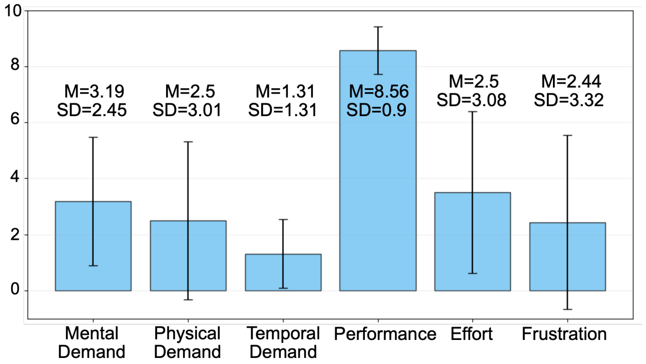


Figure 6: NASA-TLX responses from the user study. Higher scores on the *Performance* dimension indicate better outcomes, whereas lower scores on the remaining dimensions reflect better outcomes.

5.4 Perceived Cognitive Load and Learning Curve

Participants generally found TouchScribe usable and easy to learn, though they reported moderate cognitive effort and a noticeable learning curve related to gesture use and hand positioning.

Although participants rated TouchScribe as useful (M=5.5, SD=1.69), the task design imposed noticeable but moderate cognitive demands (NASA-TLX: M=3.19, SD=2.45 out of 10), as participants needed to remember descriptions and associate them with the corresponding objects. As P3 noted, *"I had to pay attention to try to remember what it was saying,"* and P6 described the experience as *"like a memory test."* Additionally, the walk-up-and-use study design introduced extra effort in learning the mappings between hand gestures and the description categories supported by TouchScribe.

On the other hand, participants generally perceived the gestures as common and easy to learn; however, the *hold+swipe-up* gesture for accessing text was considered less intuitive. As P1 noted, *"I would not guess this unless you told me the inspiration was that (from VoiceOver)"*. In addition, participants reported that positioning their hands within the camera frame required effort. P6 explained, *"In theory, it (TouchScribe) is very quick to learn, which only took us 2 minutes to go through all. In practice, it's a learning process of getting the hand placement just right, because it's a little bit finicky."*

The combined effects of these challenges, including interruptions from gesture recognition errors and delays introduced by hierarchical feedback, occasionally increased cognitive effort. However, compared to current practices, participants still perceived hand-based information access as convenient. As P1 remarked, *"If you are at the store and you have to continually find ways to read different products, using hands would be easier and more convenient."* We discuss potential improvements in gesture customization and camera aiming to reduce these demands and enhance TouchScribe's usability in Sections 7.1 and 7.3.

6 Technical Evaluation Results

Using data from the user study, we conducted a technical evaluation of (i) the hand gesture recognition performance of our pipeline in live video stream, (ii) the accuracy of the system-generated descriptions (Table 4), and (iii) the latency between gesture input and description output (Table 5).

6.1 Performance of Hand Gesture and Gesture Recognition in Live Stream

The goal of this evaluation was to assess the accuracy of our custom gesture recognition models in live video settings, beyond single-image performance. Unlike conventional model evaluations, we considered the combined performance of the recognition models and the temporal smoothing function (Section 3.5). We reviewed gesture event logs and keyframes collected during the user study.

6.1.1 Dataset and Analysis. During the user study, all keyframes and corresponding timestamps were automatically logged whenever a stable gesture state transition was detected. This enabled evaluation of both the gesture recognition models across sequences of frames and the effectiveness of the temporal smoothing algorithm. Each keyframe was labeled with a gesture state, including *hold*, *touch*, *point*, and *out of view*. In total, we collected 1,994 gesture instances, with 1,077 from the left hand and 917 from the right hand.

Each keyframe was manually annotated with a ground-truth gesture label by the research team. Gesture classes were assigned based on the visible hand pose, while the *out of view* label was used when the wrist keypoint was not visible or when fingers were partially occluded by image boundaries, objects, or the other hand, conditions under which the Google MediaPipe hand landmark model [9] may fail. We evaluated model performance by comparing predicted gestures to these ground-truth labels and computing standard metrics, including accuracy, precision, recall, F1-score, and confusion matrices for both hands.

6.1.2 Results. Among the 1,994 manually labeled instances, the model achieved an F_1-score of 0.77. Among the gesture classes (Figure 7), the *"hold"* gesture achieved the highest performance (F_1 = 0.84, precision = 0.97, recall = 0.75). The *"touch"* gesture showed high recall (0.87) but lower precision (0.60), resulting in an F_1-score of 0.71. Similarly, the *"out of view"* class achieved an F_1-score of 0.74 (precision = 0.66, recall = 0.84). The *"point"* gesture, which had the fewest instances (N = 102), showed the lowest performance (F_1 = 0.44, precision = 0.36, recall = 0.56).

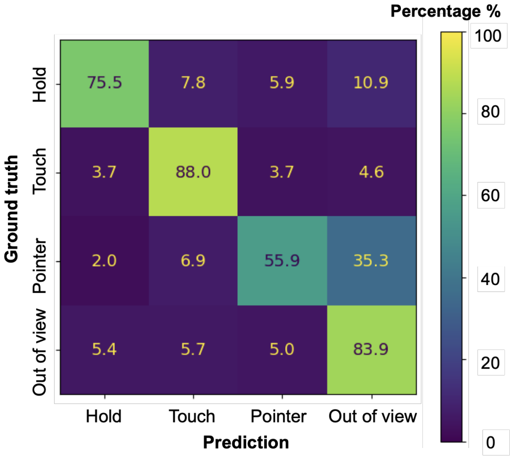


Figure 7: Confusion matrix for hand event detection of both hands.

False positives and negatives were observed under various conditions (Figure 8). For example, holding objects often resulted in partial or full hand occlusion, such as when grasping a juice carton or a box of chocolate bars (Figure 8e, f), which led to incorrect hand landmark detection and both types of errors. Similar occlusions occurred during discrete gestures like *hold+point* and *hold+swipe-up*. In addition, body movements and camera angles occasionally resulted in motion blurs and hands or fingers being partially cropped or outside the frame (Figure 8g, h). We discuss these camera-related issues and potential solutions in Section 7.3, as well as broader improvements to our vision-only approach in Section 7.2.

6.2 Latency of Delivered Descriptions

Next, we measured the latency of descriptions to quantify how long users waited before feedback was read aloud. This included hand-state feedback, brief and detailed object descriptions, comparative descriptions, color labels, and object texts. We measured end-to-end latency as the time between detection of a new gesture and the onset of the corresponding spoken description. This measurement encompassed the entire processing pipeline, including gesture recognition, retrieval of hand–object contact data and cropped images via Hands23 [36], prompt construction, response generation by VLMs (from Figure 3a to Figure 3c), and text-to-speech synthesis.

6.2.1 Dataset and Analysis. In total, we analyzed all descriptions presented to participants during the study, comprising 1,143 instances of *hand-state feedback*, 416 instances of *brief object descriptions* generated by Moondream [15], 208 instances of *detailed object descriptions* generated by GPT-4o [11], 35 instances of *comparative descriptions* generated by GPT-4o, 529 instances of *color labels*, and

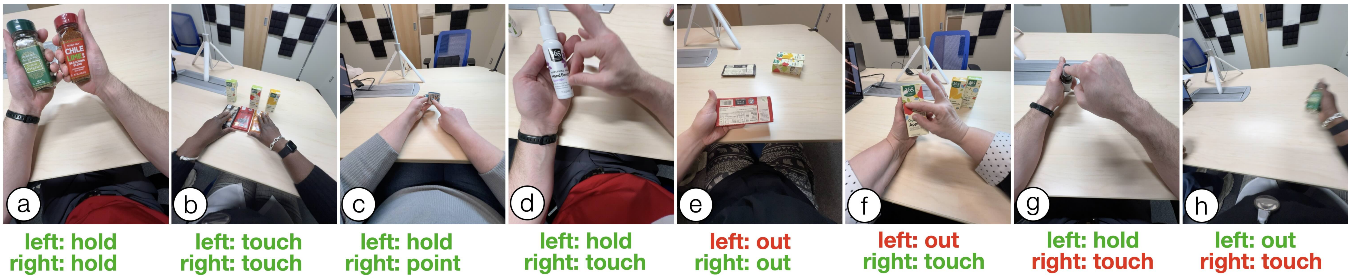


Figure 8: Example keyframes extracted by TouchScribe and the corresponding recognized gesture classes. (a–d) TouchScribe successfully identified gestures from both hands across varying camera viewpoints. (e) Recognition became challenging sometimes when hands were occluded, either by a larger object (e.g., a chocolate bar) or (f) by the user's other hand during bimanual interactions. (g) Gestures could also be misclassified under certain camera angles or hand postures, such as when the finger in a pointing gesture appears not extended from the camera's viewpoint. (h) Motion blur caused by camera or hand motion also influenced recognition reliability.

143 instances of *object texts*. We measured the latency for each description type as the time between detecting a new gesture and sending the corresponding frame to VLMs, and the moment when the resulting description was delivered to the user. Below, we first report the processing time of individual components in the pipeline, followed by the end-to-end latency experienced by users.

6.2.2 Results - latency of each model. Because the latency of each component contributes to the overall end-to-end delay, we report individual model latencies to illustrate their respective performance (Table 5). Under the hardware configuration described in Section 3.3, Hands23 exhibited an average latency of 0.87 seconds (SD=0.86), Moondream averaged 0.48 seconds (SD=0.62), and GPT-4o incurred the highest latency, with a mean of 3.07 seconds (SD=3.08).

6.2.3 Results - end-to-end latency between gesture issued to descriptions delivered. Under the hierarchical feedback design, *brief object descriptions* typically followed *hand-state feedback*, with detailed or comparative descriptions presented subsequently. In contrast, discrete gestures such as *hold+point* for color identification and *hold+swipe-up* for text retrieval allowed users to interrupt ongoing narration and quickly access targeted information (Section 3.7).

In terms of end-to-end latency, *hand-state feedback* exhibited a mean delay of 0.56 seconds (SD=0.91), providing near-immediate confirmation of system perception. Among all feedback types, color labels triggered by the *hold+point* gesture had the lowest latency (M=0.09s, SD=0.17), while object text retrieval via *hold+swipe-up* averaged 0.57 seconds (SD = 0.58).

In contrast, *brief object descriptions* generated by Moondream averaged 5.36 seconds (SD=3.42), followed by *detailed object descriptions* from GPT-4o with a mean latency of 10.3 seconds (SD=4.02). *Comparative descriptions* from GPT-4o exhibited the highest latency, averaging 14.0 seconds (SD=3.06). Notably, these latency values account for the completion of prior descriptions, during which users were engaged with ongoing audio output rather than waiting idly.

6.3 Accuracy of Object Descriptions from VLMs

Lastly, we evaluated the accuracy of descriptions generated by VLMs. The goal was to determine whether the information presented to the user is accurate and relevant. For each referenced frame, we assessed whether TouchScribe correctly described the interacted object and whether any hallucinations appeared in the generated descriptions.

6.3.1 Dataset and Analysis. In total, we collected 802 descriptions from the study, including 416 *brief object descriptions* from Moondream, 143 *object texts*, 208 *detailed object descriptions*, and 35 *comparative descriptions* generated by GPT-4o. All instances were manually annotated for correctness. Descriptions were deemed incorrect if the system misidentified the interacted object or exhibited hallucinations.

6.3.2 Results. We evaluated the accuracy of 802 descriptions. *Brief object descriptions* generated by Moondream achieved an accuracy of 91.59% (381 out of 416). *Detailed object descriptions* generated by GPT-4o reached 93.27% accuracy (194 out of 208). *Comparative descriptions* generated by GPT-4o achieved 91.43% accuracy (32 out of 35). Overall, the descriptions demonstrated strong accuracy, with common errors including misidentifying a chocolate bar as a book, hallucinating a mouse when users rested their hands on the table, referencing the table instead of the held object, or failing to describe objects when they were occluded by hands. In contrast, *object texts* achieved lower accuracy at 67.83% (97 out of 143). This was primarily due to challenges in recognizing text on the cylindrical bottles used in the study, where curvature often distorted or partially occluded the text. Participants frequently relied on trial-and-error repositioning to present readable text to the camera. We discuss potential mitigation strategies in Section 7.3.

7 Discussion and Future Work

We discuss our lessons learned and design implications for generating live object descriptions with hands as natural information cursors.

7.1 Supporting Customization and Adaptation of Broader Gesture Set

To our knowledge, TouchScribe is the first system to deliver live, rich descriptions driven by diverse hand–object interactions. TouchScribe requires users to learn a predefined gesture set. This set, though informed by prior research on *gesture nature* and *BLV familiarity* (Section 3.2), resulted in perceived cognitive load by our participants (Section 5.4). Nevertheless, qualitative feedback highlighted the utility of gestures and the enhanced sense of agency they afforded in accessing object information (Section 5.3).

The current gesture set serves as a foundation that can be extended through user customization [50, 79, 90, 96, 103]. Such flexibility is essential given the diversity of gesture preferences and contextual needs among BLV users. Social acceptability, in particular, could play a key role in shaping gesture choice. For example, mid-air gestures may be more suitable in private settings, where BLV users have been observed performing metaphoric gestures for tasks such as TV control, while sighted users tend to employ symbolic gestures [38]. In contrast, subtle micro-gestures or touchscreen interactions are often preferred in public environments due to their discreet nature [79, 96].

Accordingly, while TouchScribe currently incorporates two micro-gesture interactions (e.g., pointing to a held object for colors and swiping up for available texts) to demonstrate feasibility, the gesture vocabulary could be expanded by drawing on interaction techniques from existing assistive technologies, such as touchscreen screen readers. Examples include swiping left or right to navigate at the word level, using multi-finger swipes to access higher-level semantic information, and familiar interactions such as pinch-to-zoom for localized text exploration.

Future work could involve systematic elicitation studies with BLV users to capture gesture preferences across public and private contexts, as well as the development of adaptive AI companions capable of learning and personalizing gesture mappings over time.

7.2 Design Implications for Low-Latency, Context-Aware Gesture Recognition

Hand movements are inherently complex and dynamic, making them difficult to capture reliably using a camera stream alone. Participants observed occasional interruptions in descriptions while using TouchScribe (Section 5.3), which were attributable to limitations in our custom gesture recognition models (Section 6.1). Even when at rest, hands may unintentionally resemble supported gestures. These erroneously detected gestures prompted TouchScribe to start generating new descriptions.

Incorporating additional contextual cues could be essential to mitigate unintended gesture recognition by better distinguishing intentional from unintentional hand activities. For example, high-level user activities can be inferred from full-body posture [24] and enriched through on-body sensors or wearable devices [25, 55, 68, 77, 103], enabling the system to disregard situations in which hands are merely resting on objects or laps, or casually moving during locomotion. Furthermore, object contact and categories may be inferred from complementary sensing modalities, such as acoustic signals [52] and electromyography [39]. Incorporating diverse sensing techniques could help reduce reliance on a vision-only pipeline, particularly susceptible to occlusion, and support cross-validation of gesture and object recognition across modalities (Section 6.1).

Beyond sensor fusion, embedding common knowledge about everyday object use, typical hand postures, and users' habitual interaction patterns could further filter out irrelevant contexts, such as hands resting on tables or interacting with familiar items like keyboards, laptops, or mice. Additionally, inaccuracies in text recognition arising from curved surfaces (Section 6.3) could be alleviated by recognizing and combining texts from multiple previously captured views of an object's surface.

Recent advances in VLMs, including improvements in both accuracy and latency and the emergence of lightweight models such as Gemini-Flash [8], suggest that system responsiveness will continue to improve. Reduced latency also allows greater temporal budget for incorporating these complementary sensing components into the description pipeline, which could further enhance overall system reliability and accuracy.

7.3 Trade Offs between Camera Devices, Configurations and Practicality

Camera-based ATs face several long-standing challenges [45, 49, 61, 62, 73, 84, 94, 95], including maintaining target objects within the camera frame [45, 95], ensuring adequate coverage of essential visual content [49, 73], and addressing the social acceptability of camera setups [26, 57, 83]. While these considerations informed the design of TouchScribe (Sections 3.2 and 3.3), further work is needed to support practical deployment in real-world contexts.

In TouchScribe, we employed a neck-mounted smartphone to free users' hands and approximate an egocentric perspective. This design was inspired by the potential of emerging smart glasses, which at the time of development involved several trade-offs. Smartphones, by contrast, offered more accessible APIs than commercial smart glasses, greater flexibility in adjusting camera resolution and FoV (e.g., standard versus wide-angle), and sufficient battery life to support extended study sessions. Although this setup met our research needs, neck-mounted cameras differ from head-mounted configurations, requiring additional synchronization between head orientation and hand movements [46, 48]. Moreover, such setups may be uncomfortable for prolonged use or raise concerns regarding social acceptability in everyday contexts [26, 31, 83], especially given varying privacy sensitivities among BLV users and bystanders.

We selected a wide FoV (FoV; 120°) rather than a standard FoV (77°) to balance coverage and distortion. While standard FoV lenses introduce minimal distortion and support more reliable hand detection, their limited coverage makes it difficult to capture both hands and relevant objects simultaneously. Consequently, we opted for a wide-angle lens to increase coverage despite its greater distortion, which negatively affected hand detection performance (Section 6.1). Although TouchScribe provided feedback on perceived hand states, participants were often unaware of the camera's intrinsic limitations, as reflected in comments such as: *"I'm wondering, does closer to the camera matter?* (P2) and *"I'm blind, so I don't think about how the camera looks and stuff. So this is all good learning."* (P5).

Building on this feedback, future research could explore a broader range of camera-mounting configurations (e.g., body-mounted or head-mounted) to better accommodate individual preferences, comfort, and social contexts. This may involve integrating additional sensors, such as IMUs in wrist- or head-mounted devices, and providing feedback to address head–hand misalignment, such as haptic–audio guidance techniques for camera aiming [23, 53]. Adaptive lens-selection strategies based on hand–camera distance and object distribution could further improve coverage and accuracy; for example, switching to a wide FoV to cover multiple objects and to a standard FoV when focusing on a single object.

7.4 Gesture-driven Descriptions Beyond Physical Reach

TouchScribe delivers live visual descriptions driven by hand–object interactions within reach. Participants found this approach intuitive (Section 5.3) and more efficient than photo-capturing and chat-based interactions in current assistive apps. While TouchScribe centers on holding and touching objects, this may not always be feasible due to social stigma, safety concerns, or personal comfort. We discuss circumstances that limit tactile engagement, and outline potential ways to support gesture-based interaction even when physical reach is constrained.

Cultural taboos surrounding public touch, reinforced by norms such as the ubiquitous museum rule of "don't touch", can lead BLV individuals to internalize tactile exploration as socially inappropriate [17]. Additionally, some BLV individuals may avoid touch due to negative prior experiences, such as being compelled to explore unfamiliar objects without preparation, consent, or agency [21, 74]. Beyond social stigma, tactile exploration can also present safety concerns, especially during public health crises such as the COVID-19 pandemic [5, 6, 27]. These challenges are further compounded by physical constraints, as some objects of interest, such as items placed on high shelves in grocery stores, may be inaccessible.

To extend gesture-driven descriptions beyond direct physical reach, future systems could build upon prior work on interaction proxies [69, 71, 108] and camera motion–enabled live description tools [18, 34] (Table 1). For example, after receiving an initial overview of a visual scene and confirming interest, users could employ subtle mid-air gestures [14, 55] (e.g., pinch) or touch gestures on an interaction proxy (e.g., touchscreen) to navigate details with audio feedback. Such integrations would broaden access to visual environments both within and beyond physical reach.

8 Conclusion

In this work, we introduced TouchScribe, a system that augments hand-object interactions with automated, live visual descriptions. By leveraging egocentric hand gestures as information cursors, TouchScribe enables users to enrich their understanding of objects through diverse interaction patterns, such as holding or touching an object to receive hierarchical descriptions, comparing objects by holding them side by side, and swiping upward to read available text. Through a controlled user study and technical evaluation, we demonstrated that TouchScribe delivers reasonably accurate, timely, and informative feedback to support BLV users across a range of object exploration tasks. Participants perceived TouchScribe to be easy to learn and intuitive, and felt in control when accessing object information. Finally, we discussed implications for real-world deployment, including supporting gesture and information customization, improving gesture recognition and description accuracy through broader contextual awareness, considering diverse camera configurations and social acceptability, and extending hand-driven interaction beyond physical reach.

Acknowledgments

We thank our anonymous reviewers and all the participants in our study for their feedback and suggestions. Ruei-Che Chang was supported by the Apple Scholars in AI/ML PhD fellowship. Rosiana Natalie was partially supported by the Michigan Data Science Fellowship at the University of Michigan. This research was also supported in part by a Google Academic Research Award and a Google Cloud Platform Credit Award.

References

[1] 2025. Aira. https://aira.io/

[2] 2025. BeMyEyes. https://www.bemyeyes.com/

[3] 2025. ChatGPT. https://openai.com/chatgpt/overview/

[4] 2025. ChatGPT can now see, hear, and speak. https://openai.com/index/chatgpt-can-now-see-hear-and-speak/

[5] 2025. Coronavirus restrictions put extra burden on the blind community: Experts. https://abcnews.go.com/Health/coronavirus-restrictions-put-extra-burden-blind-community-experts/story?id=69674998

[6] 2025. COVID-19: Confessions from a Blind Germaphobe. https://www.afb.org/aw/21/4/16971

[7] 2025. Envision AI. https://www.letsenvision.com/

[8] 2025. Gemini 2.5 Flash Best for fast performance on everyday tasks. https://deepmind.google/models/gemini/flash/

[9] 2025. Google AI for Developers - Gesture recognition task guide. https://ai.google.dev/edge/mediapipe/solutions/vision/gesture_recognizer

[10] 2025. Google Android TalkBack. https://support.google.com/accessibility/android/answer/6283677?hl=en

[11] 2025. GPT-4 Omni. https://openai.com/index/hello-gpt-4o/

[12] 2025. hand-gesture-recognition-using-mediapipe. https://github.com/Kazuhito00/hand-gesture-recognition-using-mediapipe

[13] 2025. Introducing Be My AI (formerly Virtual Volunteer) for People who are Blind or Have Low Vision, Powered by OpenAI's GPT-4. https://www.bemyeyes.com/blog/introducing-be-my-eyes-virtual-volunteer

[14] 2025. Learn VoiceOver gestures with Apple Vision Pro. https://support.apple.com/guide/apple-vision-pro/learn-voiceover-gestures-tanf5c9d873d/visionos

[15] 2025. Moondream. https://moondream.ai/

[16] 2025. Orcam: Empowering Accessibility with AI. https://www.orcam.com/en-us/home?srsltid=AfmBOoqOElsZ58Z5kmCs_LDgPW7eq4WRpLVMbFGtydZn4aqayDxc7wxR

[17] 2025. Please do touch the art! https://redefine-able.thepealecenter.org/please-do-touch-the-art/

[18] 2025. SeeingAI. https://www.seeingai.com/

[19] 2025. Use VoiceOver gestures on iPhone. https://support.apple.com/guide/iphone/use-voiceover-gestures-iph3e2e2281/ios

[20] 2025. webcolors 24.11.1. https://pypi.org/project/webcolors/

[21] retrieved in 2025. Trust Through Touch: Creating Positive Interactions. https://www.tsbvi.edu/tx-senseabilities/issues/tx-senseabilities-spring-2023-issue/trust-through-touch

[22] Dustin Adams, Lourdes Morales, and Sri Kurniawan. 2013. A qualitative study to support a blind photography mobile application. In *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments* (Rhodes, Greece) *(PETRA '13)*. Association for Computing Machinery, New York, NY, USA, Article 25, 8 pages. doi:10.1145/2504335.2504360

[23] Dustin Adams, Lourdes Morales, and Sri Kurniawan. 2013. A qualitative study to support a blind photography mobile application. In *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments* (Rhodes, Greece) *(PETRA '13)*. Association for Computing Machinery, New York, NY, USA, Article 25, 8 pages. doi:10.1145/2504335.2504360

[24] Karan Ahuja, Chris Harrison, Mayank Goel, and Robert Xiao. 2019. MeCap: Whole-Body Digitization for Low-Cost VR/AR Headsets. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 453–462. doi:10.1145/3332165.3347889

[25] Karan Ahuja, Sven Mayer, Mayank Goel, and Chris Harrison. 2021. Pose-on-the-Go: Approximating User Pose with Smartphone Sensor Fusion and Inverse Kinematics. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 9, 12 pages. doi:10.1145/3411764.3445582

[26] Taslima Akter, Tousif Ahmed, Apu Kapadia, and Manohar Swaminathan. 2022. Shared Privacy Concerns of the Visually Impaired and Sighted Bystanders with Camera-Based Assistive Technologies. *ACM Trans. Access. Comput.* 15, 2, Article 11 (May 2022), 33 pages. doi:10.1145/3506857

[27] Joana Pimentel Alves, Celeste Eusébio, Maria João Carneiro, Leonor Teixeira, and Susana Mesquita. 2023. Living in an untouchable world: Barriers to recreation and tourism for Portuguese blind people during the COVID-19 pandemic. *Journal of Outdoor Recreation and Tourism* 42 (2023), 100637.

[28] Rahul Arora, Rubaiat Habib Kazi, Danny M. Kaufman, Wilmot Li, and Karan Singh. 2019. MagicalHands: Mid-Air Hand Gestures for Animating in VR. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 463–477. doi:10.1145/3332165.3347942

[29] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. 2010. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology* (New York, New York, USA) *(UIST '10)*. Association for Computing Machinery, New York, NY, USA, 333–342. doi:10.1145/1866029.1866080

[30] 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

[31] Roger Boldu, Denys J.C. Matthies, Haimo Zhang, and Suranga Nanayakkara. 2020. AiSee: An Assistive Wearable Device to Support Visually Impaired Grocery Shoppers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 119 (Dec. 2020), 25 pages. doi:10.1145/3432196

[32] Richard A Bolt. 1980. "Put-that-there" Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*. 262–270.

[33] Minjie Cai, Kris Kitani, and Yoichi Sato. 2018. Understanding hand-object manipulation by modeling the contextual relationship between actions, grasp types and object attributes. *arXiv preprint arXiv:1807.08254* (2018).

[34] Ruei-Che Chang, Yuxuan Liu, and Anhong Guo. 2024. WorldScribe: Towards Context-Aware Live Visual Descriptions. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 140, 18 pages. doi:10.1145/3654777.3676375

[35] Ruei-Che Chang, Rosiana Natalie, Wenqian Xu, Jovan Zheng Feng Yap, and Anhong Guo. 2025. Probing the Gaps in ChatGPT's Live Video Chat for Real-World Assistance for People who are Blind or Visually Impaired. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility* (Denver, Colorado, USA) *(ASSETS '25)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3663547.3746319

[36] Tianyi Cheng, Dandan Shan, Ayda Sultan Hassen, Richard Ely Locke Higgins, and David Fouhey. 2023. Towards a richer 2d understanding of hands at scale. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[37] David Costa and Carlos Duarte. 2019. Factors that Impact the Acceptability of On-Body Interaction by Users with Visual Impairments. In *Human-Computer Interaction–INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part I 17*. Springer, 267–287.

[38] Nem Khan Dim, Chaklam Silpasuwanchai, Sayan Sarcar, and Xiangshi Ren. 2016. Designing Mid-Air TV Gestures for Blind People Using User- and Choice-Based Elicitation Approaches. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (Brisbane, QLD, Australia) *(DIS '16)*. Association for Computing Machinery, New York, NY, USA, 204–214. doi:10.1145/2901790.2901834

[39] Junjun Fan, Xiangmin Fan, Feng Tian, Yang Li, Zitao Liu, Wei Sun, and Hongan Wang. 2018. What is That in Your Hand? Recognizing Grasped Objects via Forearm Electromyography Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 161 (Dec. 2018), 24 pages. doi:10.1145/3287039

[40] 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* (Lisbon, Portugal) *(ASSETS '15)*. Association for Computing Machinery, New York, NY, USA, 383–384. doi:10.1145/2700648.2811381

[41] Daniel Goldreich and Ingrid M Kanics. 2003. Tactile acuity is enhanced in blindness. *Journal of Neuroscience* 23, 8 (2003), 3439–3445.

[42] Zhitong Guan, Zeyu Xiong, and Mingming Fan. 2024. FetchAid: Making Parcel Lockers More Accessible to Blind and Low Vision People With Deep-learning Enhanced Touchscreen Guidance, Error-Recovery Mechanism, and AR-based Search Support. 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 39, 15 pages. doi:10.1145/3613904.3642213

[43] Anhong Guo, Xiang 'Anthony' Chen, Haoran Qi, Samuel White, Suman Ghosh, Chieko Asakawa, and Jeffrey P. Bigham. 2016. VizLens: A Robust and Interactive Screen Reader for Interfaces in the Real World. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 651–664. doi:10.1145/2984511.2984518

[44] Anhong Guo, Junhan Kong, Michael Rivera, Frank F. Xu, and Jeffrey P. Bigham. 2019. StateLens: A Reverse Engineering Solution for Making Existing Dynamic Touchscreens Accessible. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 371–385. doi:10.1145/3332165.3347873

[45] Anhong Guo, Saige McVea, Xu Wang, Patrick Clary, Ken Goldman, Yang Li, Yu Zhong, and Jeffrey P. Bigham. 2018. Investigating Cursor-based Interactions to Support Non-Visual Exploration in the Real World. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (Galway, Ireland) *(ASSETS '18)*. Association for Computing Machinery, New York, NY, USA, 3–14. doi:10.1145/3234695.3236339

[46] Yangha Han, Mahya Beheshti, Blake Jones, Todd E Hudson, William H Seiple, and John-Ross Rizzo. 2024. Wearables for persons with blindness and low vision: form factor matters. *Assistive Technology* 36, 1 (2024), 60–63.

[47] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[48] Marion Hersh. 2022. Wearable travel aids for blind and partially sighted people: A review with a focus on design issues. *Sensors* 22, 14 (2022), 5454.

[49] Naoki Hirabayashi, Masakazu Iwamura, Zheng Cheng, Kazunori Minatani, and Koichi Kise. 2023. VisPhoto: Photography for People with Visual Impairments via Post-Production of Omnidirectional Camera Imaging. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA) *(ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, Article 6, 17 pages. doi:10.1145/3597638.3608422

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

[51] Zhanpeng Huang, Weikai Li, and Pan Hui. 2015. Ubii: Towards Seamless Interaction between Digital and Physical Worlds. In *Proceedings of the 23rd ACM International Conference on Multimedia* (Brisbane, Australia) *(MM '15)*. Association for Computing Machinery, New York, NY, USA, 341–350. doi:10.1145/2733373.2806266

[52] Yasha Iravantchi, Yi Zhao, Kenrick Kin, and Alanson P. Sample. 2023. SAWSense: Using Surface Acoustic Waves for Surface-bound Event Recognition. 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 422, 18 pages. doi:10.1145/3544548.3580991

[53] Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P. Bigham. 2011. Supporting blind photography. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility* (Dundee, Scotland, UK) *(ASSETS '11)*. Association for Computing Machinery, New York, NY, USA, 203–210. doi:10.1145/2049536.2049573

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

[55] Prerna Khanna, IV Ramakrishnan, Shubham Jain, Xiaojun Bi, and Aruna Balasubramanian. 2024. Hand Gesture Recognition for Blind Users by Tracking 3D Gesture Trajectory. 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 405, 15 pages. doi:10.1145/3613904.3642602

[56] Taejun Kim, Amy Karlson, Aakar Gupta, Tovi Grossman, Jason Wu, Parastoo Abtahi, Christopher Collins, Michael Glueck, and Hemant Bhaskar Surale. 2023. STAR: Smartphone-analogous Typing in Augmented Reality. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 116, 13 pages. doi:10.1145/3586183.3606803

[57] Marion Koelle, Torben Wallbaum, Wilko Heuten, and Susanne Boll. 2019. Evaluating a Wearable Camera's Social Acceptability In-the-Wild. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3290607.3312837

[58] Sinisa Kolaric, Alberto Raposo, and Marcelo Gattass. 2008. Direct 3D manipulation using vision-based recognition of uninstrumented hands. In *X Symposium on Virtual and Augmented Reality*. Citeseer, 212–220.

[59] Jaewook Lee, Jaylin Herskovitz, Yi-Hao Peng, and Anhong Guo. 2022. ImageExplorer: Multi-Layered Touch Exploration to Encourage Skepticism Towards Imperfect AI-Generated Image Captions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>) *(CHI*

'22). Association for Computing Machinery, New York, NY, USA, Article 462, 15 pages. doi:10.1145/3491102.3501966

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

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

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

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

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

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

[66] Minkyung Lee, Richard Green, and Mark Billinghurst. 2008. 3D natural hand interaction for AR applications. In *2008 23rd International Conference Image and Vision Computing New Zealand*. IEEE, 1–6.

[67] Fabrizio Leo, Monica Gori, and Alessandra Sciutti. 2022. Early blindness modulates haptic object recognition. *Frontiers in Human Neuroscience* 16 (2022), 941593.

[68] Yunzhi Li, Vimal Mollyn, Kuang Yuan, and Patrick Carrington. 2024. WheelPoser: Sparse-IMU Based Body Pose Estimation for Wheelchair Users. 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 8, 17 pages. doi:10.1145/3663548.3675638

[69] Chen Liang, Yuxuan Liu, Martez Mott, and Anhong Guo. 2025. HandProxy: Expanding the Affordances of Speech Interfaces in Immersive Environments with a Virtual Proxy Hand. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3, Article 107 (Sept. 2025), 30 pages. doi:10.1145/3749484

[70] Jingyuan Liu, Hongbo Fu, and Chiew-Lan Tai. 2020. PoseTween: Pose-driven Tween Animation. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 791–804. doi:10.1145/3379337.3415822

[71] Tao Lu, Hongxiao Zheng, Tianying Zhang, Xuhai "Orson" Xu, and Anhong Guo. 2024. InteractOut: Leveraging Interaction Proxies as Input Manipulation Strategies for Reducing Smartphone Overuse. 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 245, 19 pages. doi:10.1145/3613904.3642317

[72] Minghuang Ma, Haoqi Fan, and Kris M Kitani. 2016. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1894–1903.

[73] Roberto Manduchi and James M. Coughlan. 2014. The last meter: blind visual guidance to a target. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3113–3122. doi:10.1145/2556288.2557328

[74] Mike McLinden and Steve Mccall. 2016. *Learning through touch: Supporting children with visual impairments and additional difficulties*. David Fulton Publishers.

[75] Alexander J. Medeiros, Lee Stearns, Leah Findlater, Chuan Chen, and Jon E. Froehlich. 2017. Recognizing Clothing Colors and Visual Textures Using a Finger-Mounted Camera: An Initial Investigation. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery, New York, NY, USA, 393–394. doi:10.1145/3132525.3134805

[76] Daniel Mendes, Fernando Fonseca, Bruno Araujo, Alfredo Ferreira, and Joaquim Jorge. 2014. Mid-air interactions above stereoscopic interactive tables. In *2014 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, 3–10.

[77] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 2023. IMUPoser: Full-Body Pose Estimation using IMUs in Phones, Watches, and Earbuds. 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 529, 12 pages. doi:10.1145/3544548.3581392

[78] 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

[79] Uran Oh and Leah Findlater. 2013. The challenges and potential of end-user gesture customization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 1129–1138. doi:10.1145/2470654.2466145

[80] Uran Oh and Leah Findlater. 2014. Design of and subjective response to on-body input for people with visual impairments. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility* (Rochester, New York, USA) *(ASSETS '14)*. Association for Computing Machinery, New York, NY, USA, 115–122. doi:10.1145/2661334.2661376

[81] Uran Oh, Lee Stearns, Alisha Pradhan, Jon E. Froehlich, and Leah Findlater. 2017. Investigating Microinteractions for People with Visual Impairments and the Potential Role of On-Body Interaction. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery, New York, NY, USA, 22–31. doi:10.1145/3132525.3132536

[82] SK Ong and ZB Wang. 2011. Augmented assembly technologies based on 3D bare-hand interaction. *CIRP annals* 60, 1 (2011), 1–4.

[83] Halley Profita, Reem Albaghli, Leah Findlater, Paul Jaeger, and Shaun K. Kane. 2016. The AT Effect: How Disability Affects the Perceived Social Acceptability of Head-Mounted Display Use. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4884–4895. doi:10.1145/2858036.2858130

[84] Manaswi Saha, Alexander J. Fiannaca, Melanie Kneisel, Edward Cutrell, and Meredith Ringel Morris. 2019. Closing the Gap: Designing for the Last-Few-Meters Wayfinding Problem for People with Visual Impairments. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 222–235. doi:10.1145/3308561.3353776

[85] Tsubasa Saito and Takashi Ijiri. 2021. Animating Various Characters using Arm Gestures in Virtual Reality Environment. In *Adjunct Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 29–31. doi:10.1145/3474349.3480220

[86] Lee Stearns, Victor DeSouza, Jessica Yin, Leah Findlater, and Jon E. Froehlich. 2017. Augmented Reality Magnification for Low Vision Users with the Microsoft Hololens and a Finger-Worn Camera. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery, New York, NY, USA, 361–362. doi:10.1145/3132525.3134812

[87] 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 Trans. Access. Comput.* 9, 1, Article 1 (Oct. 2016), 38 pages. doi:10.1145/2914793

[88] Lee Stearns, Ruofei Du, Uran Oh, Yumeng Wang, Leah Findlater, Rama Chellappa, and Jon E Froehlich. 2015. The design and preliminary evaluation of a finger-mounted camera and feedback system to enable reading of printed text for the blind. In *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part III 13*. Springer, 615–631.

[89] Lee Stearns, Leah Findlater, and Jon E. Froehlich. 2018. Applying Transfer Learning to Recognize Clothing Patterns Using a Finger-Mounted Camera. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (Galway, Ireland) *(ASSETS '18)*. Association for Computing Machinery, New York, NY, USA, 349–351. doi:10.1145/3234695.3241015

[90] Lee Stearns, Uran Oh, Leah Findlater, and Jon E. Froehlich. 2018. TouchCam: Realtime Recognition of Location-Specific On-Body Gestures to Support Users with Visual Impairments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 164 (Jan. 2018), 23 pages. doi:10.1145/3161416

[91] Hemant Bhaskar Surale, Fabrice Matulic, and Daniel Vogel. 2019. Experimental Analysis of Barehand Mid-air Mode-Switching Techniques in Virtual Reality. 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, 1–14. doi:10.1145/3290605.3300426

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

[93] Marynel Vázquez and Aaron Steinfeld. 2012. Helping visually impaired users properly aim a camera. In *Proceedings of the 14th International ACM SIGACCESS*

Conference on Computers and Accessibility (Boulder, Colorado, USA) *(ASSETS '12)*. Association for Computing Machinery, New York, NY, USA, 95–102. doi:10.1145/2384916.2384934

[94] Marynel Vázquez and Aaron Steinfeld. 2012. Helping visually impaired users properly aim a camera. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility* (Boulder, Colorado, USA) *(ASSETS '12)*. Association for Computing Machinery, New York, NY, USA, 95–102. doi:10.1145/2384916.2384934

[95] Marynel Vázquez and Aaron Steinfeld. 2012. Helping visually impaired users properly aim a camera. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility* (Boulder, Colorado, USA) *(ASSETS '12)*. Association for Computing Machinery, New York, NY, USA, 95–102. doi:10.1145/2384916.2384934

[96] Ruolin Wang, Chun Yu, Xing-Dong Yang, Weijie He, and Yuanchun Shi. 2019. EarTouch: Facilitating Smartphone Use for Visually Impaired People in Mobile and Public Scenarios. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300254

[97] Tianyi Wang, Xun Qian, Fengming He, Xiyun Hu, Yuanzhi Cao, and Karthik Ramani. 2021. GesturAR: An Authoring System for Creating Freehand Interactive Augmented Reality Applications. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 552–567. doi:10.1145/3472749.3474769

[98] Maarten WA Wijntjes, Thijs Van Lienen, Ilse M Verstijnen, and Astrid ML Kappers. 2008. The influence of picture size on recognition and exploratory behaviour in raised-line drawings. *Perception* 37, 4 (2008), 602–614.

[99] Ans Withagen, Astrid ML Kappers, Mathijs PJ Vervloed, Harry Knoors, and Ludo Verhoeven. 2012. Haptic object matching by blind and sighted adults and children. *Acta Psychologica* 139, 2 (2012), 261–271.

[100] Ans Withagen, Mathijs PJ Vervloed, Neeltje M Janssen, Harry Knoors, and Ludo Verhoeven. 2010. Tactile functioning in children who are blind: A clinical perspective. *Journal of Visual Impairment & Blindness* 104, 1 (2010), 43–54.

[101] Jingyi Xie, Rui Yu, He Zhang, Syed Masum Billah, Sooyeon Lee, and John M. Carroll. 2025. Beyond Visual Perception: Insights from Smartphone Interaction of Visually Impaired Users with Large Multimodal Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 62, 17 pages. doi:10.1145/3706598.3714210

[102] Andi Xu, Minyu Cai, Dier Hou, Ruei-Che Chang, and Anhong Guo. 2024. ImageExplorer Deployment: Understanding Text-Based and Touch-Based Image Exploration in the Wild. In *Proceedings of the 21st International Web for All Conference (W4A '24)*. Association for Computing Machinery, New York, NY, USA, 59–69. doi:10.1145/3677846.3677861

[103] Xuhai Xu, Jun Gong, Carolina Brum, Lilian Liang, Bongsoo Suh, Shivam Kumar Gupta, Yash Agarwal, Laurence Lindsey, Runchang Kang, Behrooz Shahsavari, Tu Nguyen, Heriberto Nieto, Scott E Hudson, Charlie Maalouf, Jax Seyed Mousavi, and Gierad Laput. 2022. Enabling Hand Gesture Customization on Wrist-Worn Devices. 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 496, 19 pages. doi:10.1145/3491102.3501904

[104] Xing-Dong Yang, Tovi Grossman, Daniel Wigdor, and George Fitzmaurice. 2012. Magic finger: always-available input through finger instrumentation. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (Cambridge, Massachusetts, USA) *(UIST '12)*. Association for Computing Machinery, New York, NY, USA, 147–156. doi:10.1145/2380116.2380137

[105] Xin Yi, Chun Yu, Mingrui Zhang, Sida Gao, Ke Sun, and Yuanchun Shi. 2015. ATK: Enabling Ten-Finger Freehand Typing in Air Based on 3D Hand Tracking Data. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 539–548. doi:10.1145/2807442.2807504

[106] Shahrouz Yousefi, Mhretab Kidane, Yeray Delgado, Julio Chana, and Nico Reski. 2016. 3D gesture-based interaction for immersive experience in mobile VR. In *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2121–2126.

[107] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343* (2023).

[108] Xiaoyi Zhang, Anne Spencer Ross, Anat Caspi, James Fogarty, and Jacob O. Wobbrock. 2017. Interaction Proxies for Runtime Repair and Enhancement of Mobile Application Accessibility. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6024–6037. doi:10.1145/3025453.3025846

[109] 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 27, 12 pages. doi:10.1145/3411764.3445578

[110] Qian Zhou, David Ledo, George Fitzmaurice, and Fraser Anderson. 2024. TimeTunnel: Integrating Spatial and Temporal Motion Editing for Character Animation in Virtual Reality. 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 101, 17 pages. doi:10.1145/3613904.3641927

[111] Zhongyi Zhou and Koji Yatani. 2022. Gesture-aware Interactive Machine Teaching with In-situ Object Annotations. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 27, 14 pages. doi:10.1145/3526113.3545648

A Tables

Table 3: Participant demographic information, referred to as P1 to P8.

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

Table 4: Accuracy of object descriptions generated by VLMs.

Description Type	Instances	Correct	Accuracy (%)
Object labels (Moondream)	416	381	91.59
Detailed descriptions (GPT-4o)	208	194	93.27
Comparative descriptions (GPT-4o)	35	32	91.43
Object texts (GPT-4o)	143	97	67.83

Table 5: Latency results for description generation, reported as mean and standard deviation (SD) in seconds.

Description Type / Model	Instances	Avg. Latency (s)	SD (s)
Model Processing Latency			
Hands23 [36]	–	0.87	0.86
Moondream [15]	–	0.48	0.62
GPT-4o [11]	–	3.07	3.08
End-to-End Latency (Gesture Identified ⟶ Description Spoken)			
Hand state descriptions	1143	0.561	0.91
Object labels (Moondream)	416	5.36	3.42
Detailed descriptions (GPT-4o)	208	10.3	4.02
Comparative descriptions (GPT-4o)	35	14.0	3.06
Object texts (GPT-4o)	143	0.57	0.58
Color labels	529	0.087	0.169