Say It My Way: Exploring Control in Conversational Visual Question Answering with Blind Users

Farnaz Zamiri Zeraati*
Department of Computer Science
University of Maryland
College Park, Maryland, USA
farnaz@umd.edu

Yang Cao*
University of Texas at Austin
Austin, Texas, USA
ycao55@utexas.edu

Yuehan Qiao
College of Information
University of Maryland, College Park
College Park, Maryland, USA
yhqiao@umd.edu

Hal Daumé III
University of Maryland
College Park, Maryland, USA
hal@umiacs.umd.edu

Hernisa Kacorri
College of Information
University of Maryland
College Park, Maryland, USA
hernisa@umd.edu

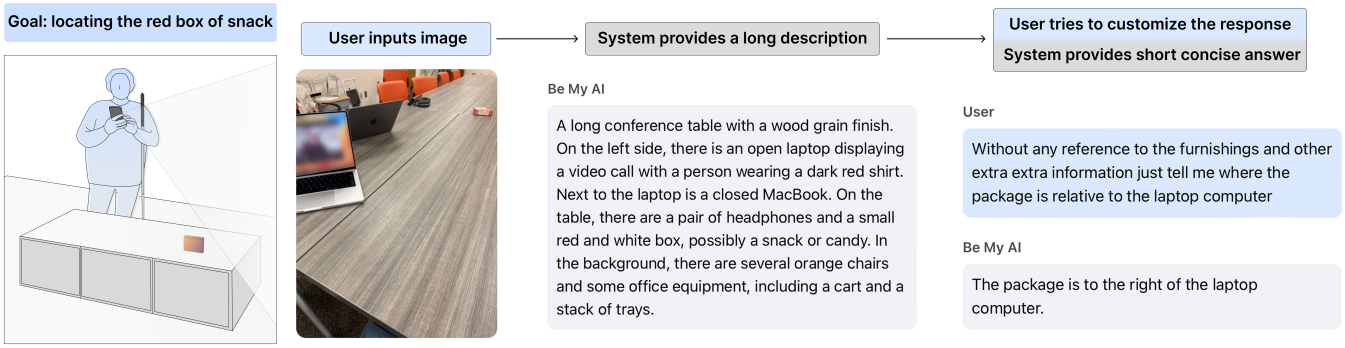


Figure 1: During an in-lab scenario, P11 experimented with stylistic customization to control the level of detail in the system's description and focus the response on locating the snack box.

Abstract

Prompting and steering techniques are well established in general-purpose generative AI, yet assistive visual question answering (VQA) tools for blind users still follow rigid interaction patterns with limited opportunities for customization. User control can be helpful when system responses are misaligned with their goals and contexts, a gap that becomes especially consequential for blind users that may rely on these systems for access. We invite 11 blind users to customize their interactions with a real-world conversational VQA system. Drawing on 418 interactions, reflections, and post-study interviews, we analyze prompting-based techniques participants adopted, including those introduced in the study and those developed independently in real-world settings. VQA interactions were often lengthy: participants averaged 3 turns, sometimes up to 21, with input text typically tenfold shorter than the responses they heard. Built on state-of-the-art LLMs, the system lacked verbosity controls, was limited in estimating distance in space and time, relied on inaccessible image framing, and offered little to no camera guidance. We discuss how customization techniques such as prompt engineering can help participants work around these limitations. Alongside a new publicly available dataset, we offer insights for interaction design at both query and system levels.

CCS Concepts

• **Human-centered computing** → **Empirical studies in accessibility**.

Keywords

Blind users, Generative AI, Visual question answering (VQA), Personalization, Customization, Prompting strategies

*Both authors contributed equally to this research.

This work is licensed under a Creative Commons Attribution 4.0 International License.
CHI '26, Barcelona, Spain
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791834

ACM Reference Format:
Farnaz Zamiri Zeraati, Yang Cao, Yuehan Qiao, Hal Daumé III, and Hernisa Kacorri. 2026. Say It My Way: Exploring Control in Conversational Visual Question Answering with Blind Users. 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, 22 pages. https://doi.org/10.1145/3772318.3791834

1 Introduction

Visual question answering (VQA) aims to answer open-ended questions about images using natural language responses and has long been explored as a way to help blind people access visual surroundings and content. Early approaches relied on sighted assistance, such as crowd-sourced and real-time services, who could help localize objects and describe scenes [5, 14, 15, 21, 30, 58]. Advances in computer vision and natural language processing have shifted many of these approaches to automated systems that generate descriptions and answer visual questions, promising greater privacy and independence [34, 59, 69, 82, 85]. More recent developments in generative AI and vision–language models (VLMs) have enabled conversational, multi-turn interpretations of visual scenes, powering real-world applications such as the *Be My AI* feature in *Be My Eyes* [13, 30].

Yet relatively little attention has been given to how blind users can shape or control these increasingly capable systems. Existing applications typically follow a fixed interaction pattern: a user provides an image, receives a lengthy system-generated description, and may then ask follow-up questions. While functional, this design limits the degree of control blind users can exercise and the extent to which responses align with their goals, preferences, and contexts. This gap is especially important because blind users may rely on these systems in situations where efficiency and clarity are critical [37]. A lengthy or poorly targeted response can create confusion, delay action, or even pose safety risks [46]. This is in contrast to general-purpose generative AI applications, where prior work documents a wide range of approaches, such as prompt engineering [17, 32, 92] and fine-tuning models with personal data [91, 105], that let users influence the type, tone, and depth of information they receive.

Motivated by this gap, we investigate the following research question: *What are the opportunities and challenges for blind users to exert control over VQA responses?* We conduct a three-phase study examining how blind users interact with existing customization techniques for controlling AI-generated responses in conversational VQA systems. The first phase is a controlled in-lab session, where participants (N=11) are introduced to and practice a set of customization techniques, such as binary feedback, zero-shot style prompting, and chain-of-thought prompting, through predefined tasks (see Figure 1 for an example). The second is a diary study. For 10 days, participants use the system independently in their daily life, documenting naturally occurring interactions across diverse tasks and environments. In the last phase, participants join a remote semistructured interview, where they share deeper insights on their interactions, preferences, and reflections on system behavior and control. This combined approach of short in person sessions with extended multi day engagement aims to capture both immediate, guided user behaviors and longitudinal, naturally occurring interactions where customization techniques may emerge, get adapted, and evolve over time. With the reflections we aim to understand how VQA systems can better support user control beyond the prompt engineering paradigm.

Our study reveals that interactions are often lengthy and imbalanced: participant inputs are typically concise, while system responses are more than ten times longer. Conversations commonly stretch over several turns (median = 3, up to 21), especially when participants make multiple attempts to capture usable images, clarify ambiguous outputs, or pursue curiosity-driven questions. While most interactions are successful, failures arise from inaccessible or missing information in the image, system limitations, or model errors. Participants actively experiment with ways to shape responses. Many adopt customization techniques introduced in the lab—such as binary feedback and targeted prompting—while also developing new strategies in the diary study. Customization is most common in professional contexts but also appears in familiar and unfamiliar environments, with slightly higher rates in familiar settings. Through these practices, participants seek to manage verbosity, direct attention to task-relevant details, and strengthen trust in system output. Reflections across tasks highlight further needs: toggling verbosity based on time pressure, spatial descriptions grounded in user-centered reference points rather than image frames, proactive support for image capture, and verification mechanisms for high-stakes situations.

This work makes two contributions: (1) empirical findings showing how blind users adapt to current conversational VQA constraints while highlighting design opportunities for greater control to support efficiency, clarity, and trust, and (2) the Ask&Prompt dataset, which contains images, text conversations from our study with the *Be My AI* application, and contextual annotations from our research team grounded on participant feedback.

2 Related Work

2.1 Personalization in Generative AI

Personalization of large language models (LLMs) has been extensively studied across a range of natural language processing tasks, including dialogue systems [61, 67], summarization [99], and search and information retrieval [3, 107]. As commercial and publicly accessible LLMs become more widespread, personalization is increasingly viewed as a key solution to several emerging challenges–enabling more usable systems across diverse tasks [75, 100], improving alignment with individual user values [16, 35], and fostering greater user autonomy [56, 57]. In response, researchers have proposed a range of approaches to enable personalization at various stages of the machine learning pipeline. During pretraining and fine-tuning, datasets have been proposed to better represent diverse user groups, tasks, and values to ensure broader inclusivity in foundational model behavior [64, 91, 105]. Techniques such as supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) [75] have also been developed to steer models toward more desirable, value-aligned outputs [22, 73, 93, 102, 108]. In the post-training phase, studies have investigated the effectiveness of few-shot prompting and in-context learning to adapt model behavior without requiring parameter updates–allowing users to shape outputs using lightweight, example-based techniques [50, 63, 83, 104]. Building on this, recent work investigates user-facing customization at inference time, including interactive tools that support prompt exploration and output refinement through mixed-initiative designs, visual aids, and LLM-powered suggestions [17, 32, 92]. These systems have been shown to reduce cognitive load, accelerate creative workflows, and improve alignment between prompts and system responses [17, 39]. Additionally, newer interfaces move beyond text, incorporating input modalities such as annotations, scribbles, and visual elements to enhance accessibility and user control [10, 29, 76].

Extending beyond text, personalization also was explored with visual language models (VLMs) and multimodal systems [6, 72, 91]. In this study, we focus on the VLM-based visual question answering system to explore how blind users take control and customize the systems to their needs and preferences.

2.2 Blind Users and Generative AI

There is a rich literature on leveraging AI to benefit the blind community. Surveys highlight applications across text access, perception and description, and mobility, delivered through smartphones, wearables, and head-mounted cameras e.g, [26, 36, 81]. While these efforts show steady progress, reviews point out that research has focused heavily on description tasks [36, 90], and that concerns around evaluation, privacy, and safety remain [4, 44].

Researchers and developers have been quick in applying generative AI to assist blind users. VLMs such as *Be My AI* [30] among others turn photos into conversational descriptions and are being integrated into mobile apps [70] and wearables [9, 11, 12] for real-time access to visual scenes, objects, and text [8, 41]. For navigation, prototypes combine LLM reasoning with on-device sensing to offer step-by-step guidance and hazard detection [12, 55, 86, 101]. Evaluations highlight persistent challenges such as hallucinations, limited contextual awareness, and difficulty supporting goal-directed interaction [45, 54]. User studies show both enthusiasm and hesitation: blind users experiment with these tools in household, social, and safety contexts, but often develop their own verification strategies to manage uncertainty [2, 37, 79, 98]. Together, this literature points to opportunities for more reliable and personalized multimodal solutions.

Within this growing space, personalization for assistive generative AI has been explored through two complementary directions: interaction and interface design, and model-level adaptation. Examples of personalization through interaction and interface design are systems that adjust descriptions based on user intent, scene context, or verbosity [23, 68]. Closer to model-level adaptation, projects on teachable object recognition have shown how users can train systems to handle personally relevant items [53, 87, 95]. Our study builds on the first direction by examining prompting and customization techniques across both controlled lab tasks and real-world diary studies, exploring a broader view of how blind users tailor VLM-based conversational VQA system to their needs.

3 Methods

To explore how blind users interact with AI-powered conversational VQA systems and attempt to control system responses, we conduct a multi-day user study (Figure 2). Participants first visit our lab for a semi-structured interview, an introduction to the commercial VQA app *Be My AI*, and 5 carefully engineered interaction scenarios. They then complete a diary study, using *Be My AI* in daily life and sharing at least 25 interaction scenarios with their goals and reflections. Last, participants join a remote semi-structured interview to reflect on their diary entries.

3.1 Recruiting Participants

We recruited participants through three channels: our lab's mailing list, the National Federation of the Blind (NFB) [1], and word of mouth. The mailing list included blind individuals who had previously participated in our research and consented to future contact. Those in the list and the NFB received an email with a study description and consent form. Interested individuals contacted us by email and were screened for eligibility. To participate, individuals had to be 18 or older and identify as blind. All participants provided informed consent and were reminded that participation is voluntary and can be withdrawn at any time, with compensation provided for completed portions. A total of 13 participants enrolled; two withdrew after the in-lab session (one citing time conflicts), leaving 11 for analysis. Of those, four identified as totally blind and seven as legally blind (Table 1). The average age was 42 years (SD = 11.33). Nine participants identified as women and two as men. Education levels ranged from high school to a JD degree; employed participants worked in management, technology, customer service, or law. Most participants reported being somewhat familiar with AI, with a broad understanding of what it does, and none reported extensive knowledge. Most participants had used *Be My Eyes* (except P5), though fewer than half mentioned *Be My AI*. Many also used *Seeing AI* and *Aira*. A few had tried *Meta Glasses*, and one mentioned *ChatGPT*. Participants received a fixed compensation of $300, based on an estimated 15 hours of study activities at a $20/hour, including 2.5 hours for the in-lab session, 25 minutes of sharing and 35 minutes of reflection per diary day, and 2.5 hours for the final interview. This rate reflected expected effort rather than incentivizing excessive engagement, and no participant exceeded the allotted time for the in-lab or remote interviews we could track. The study protocol was approved by our institution's review board.

3.2 Introducing Customization in the Lab

During the lab visit, participants were introduced to the conversational VQA system used throughout the study: *Be My AI*, a feature of the *BeMyEyes* mobile application[1]. This phase aimed to (1) ensure participants were comfortable navigating the app and (2) expose them to existing customization techniques while gathering initial perspectives. We selected *Be My AI* because, at the time, it was the only mature, widely available, and screen-reader accessible AI-powered visual assistance tool. In-lab sessions were recorded with two devices: a laptop camera capturing the front view and a pair of smart glasses capturing the researcher's view of the environment and participant interactions.

During the session, we first collected participants' demographics and their attitudes toward and experience with technology. This was followed by a guided practice session where participants engaged with the system across five scenarios simulating real-world tasks. These scenarios were informed by frequent use cases from the VizWiz dataset [40], prior work with blind users [e.g., 15, 52, 89, 95], and discussions with a blind researcher.

We introduced a set of customization techniques, appropriate for these scenarios, grounded in established prompting techniques for large language models. Drawing on the taxonomy by Schulhoff et al. [84], we selected strategies relevant to visual assistance and adapted them to fit short, mobile-based interactions and existing functionalities in *Be My AI*. A full list of introduced techniques

[1]We designed the study using *Be My AI* v6.0. During the study, participants used v6.0-6.3, which differed only in minor bug fixes and language support.

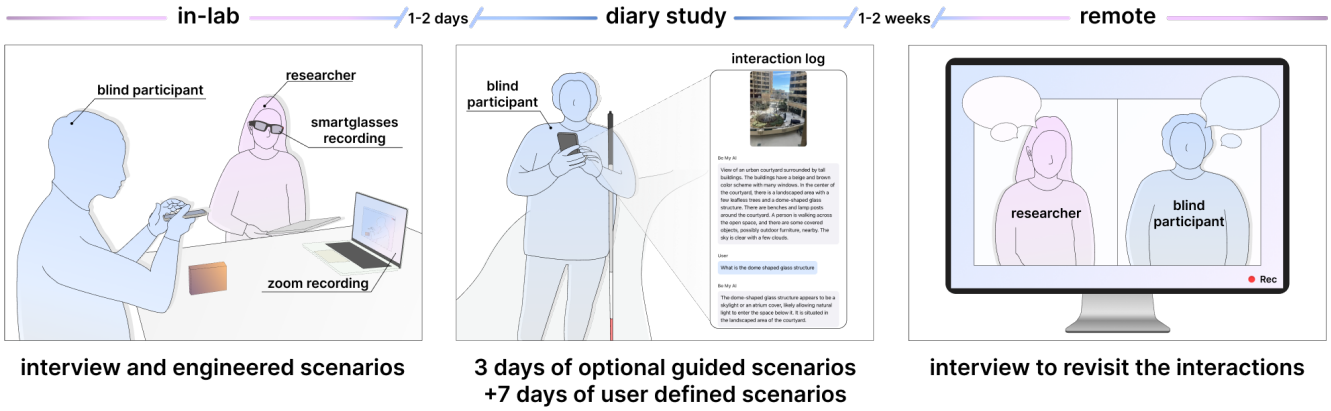


Figure 2: Our multi-day study includes an in-lab interview with scenarios, a 10-day diary, and a remote interview.

Table 1: Self-reported participant demographics, familiarity with AI, and related assistive technology use.

PID	Blindness	Age	Gender	Education	Occupation	AI Fam*	Related Assistive Tech
P1	total	30	W	associate	full-time parent	1	Be My Eyes
P2	legal	39	W	high school	customer service representative	3	Seeing AI; Be My Eyes; Aira
P3	total	52	W	some college	not employed	3	Be My Eyes; Seeing AI
P4	total	38	W	high school	not employed	2	Be My Eyes
P5	legal	53	W	master	not employed	3	-
P6	legal	31	W	some college	assistant technology tutor	3	Be My AI; Aira; Meta Glasses
P7	legal	34	M	master	budget manager	3	Be My AI; Be My Eyes; Seeing AI
P8	legal	42	W	master	management specialist	2	Be My Eyes; Seeing AI; Alexa
P9	legal	25	W	bachelor	accessibility program manager	3	Be My Eyes; Aira; ChatGPT; Seeing AI
P10	legal	56	W	master	manager of a care program	3	Seeing AI; Be My AI; Aira
P11	total	57	M	juris doctor	lawyer	3	Aira; Be My AI; Meta Glasses

*Familiarity with AI was reported in a 4-point scale: 1 = not familiar (never heard of it), 2 = slightly familiar (heard of it but don't know what it does), 3 = somewhat familiar (broad understanding of what it is/does), 4 = extremely familiar (extensive knowledge).

appears in Table 2, with detailed scenarios in Appendix A. To our knowledge no prior work has systematically documented prompting techniques for accessibility. We present our adaptation as an initial step toward such a framework. However, because *Be My AI* does not support persistent customization, participants did not experience lasting effects. To manage expectations, we informed them of this limitation but encouraged them to use the techniques as if preferences could carry over. After each scenario, participants reflected on goal achievement, ideal responses, and customization techniques, comparing them and noting preferences across contexts.

3.3 Collecting Diary Data in Everyday Contexts

A day or two after the in-lab session, participants began a 10-day diary study using *Be My AI* independently in their everyday environments. The first 3 days were lightly structured to help participants become comfortable with the system and to encourage early exploration across contexts. From Day 4 through Day 10 there was no prompting beyond submitting at least 25 interactions, experimenting with different ways of guiding the system including the customization techniques introduced in the lab. Instructions were sent via email or message, according to preference. For each interaction, participants shared a brief description and link, noting the task goal, environment familiarity, and time constraints. To support deeper reflection, they also completed a short set of end-of-day questions. We note that while our data collection focused solely on interactions submitted through *Be My AI*, participants were not restricted in their technology use and were free to use any other assistive tools throughout the diary study. Full protocol details for Days 1-3 and reflection questions appear in Appendix B and Appendix C.

3.4 Conducting Post-Diary Interviews

Within two weeks of the diary study, participants completed a virtual interview on their preferred platform (Zoom or Microsoft Teams). The goal was to gain deeper insights into their interactions, preferences, and reflections on system behavior and control. With consent, interviews were audio-recorded and transcribed.

The interview began with scenario confirmations. We selected submitted diary interactions that lacked context—such as goals, time sensitivity, familiarity with the setting or object, or task outcome—and asked participants to confirm or correct our interpretation. We also asked whether any interaction raised privacy concerns and

Table 2: Definitions and examples of the customization techniques introduced in the lab.

Technique	Definition in GenAI	Definition in our use case	Example used
Binary Feedback	Feedback that indicate positive or negative rewards [75].	Users provide a simple positive or negative feedback with the response	*"Helpful/Not helpful."*
Zero-shot: Style Prompting	Prompt to specify desired stylistic attributes of GenAI output (e.g., tone and genre) [66].	Users prompt to specify desired style to adjust following responses.	*"More/Less detail."*
Zero-shot: Intention Prompting*	Emotion prompt that incorporates phrases of psychological relevance to humans [62].	Users express their goals to guide systems tailor responses based on the task context.	*"I want to gain muscle. Is this a good option of food?"*
CoT Prompting	Offer reasoning behind to guide GenAI systems [78, 94].	Users prompt with reasoning for desirable or undesirable answers.	*"I only care about finding the soup box, so do not mention about other items."*
Image-as-text Prompting	Generate a textual description of image(s), enabling easy inclusion into a text-based prompt[42].	Users prompt with visual characteristics of objects or scenes, allowing systems to complete tasks with additional information.	*"Find my keys with a keychain that is green, round, and with raised letters."*

*Zero-shot: Intention prompting is generalized from emotion prompting.

obtained explicit consent for those they agreed to share for future research and publications.

Next, we reviewed five to six interactions per participant, selected to represent diverse tasks. For each, participants described their ideal system response in terms of content (e.g., type of information) and style (e.g., tone, level of detail), and whether their preferences would differ across contexts, such as when they were in a hurry or not.

We then explored customization techniques. Participants reflected on why they used a technique and how effective it felt, or if they had not, why not and what alternatives they could imagine. Discussions revealed not only how they customized the system, but also which techniques felt intuitive, accessible, or worth the effort in other contexts.

Finally, we gathered reflections on overall experience. Participants rated two usability statements—ease of use and comfort—on a 5-point Likert scale and answered open-ended questions about customization techniques, including which they preferred and which felt challenging or unnatural. They also reflected on future use: when they might trust the system without customization and when they would invest effort to guide or verify responses.

3.5 Analyzing Data Across Sessions

To better understand how blind participants interacted with the system and adopted customization techniques, we triangulated data from multiple sources: recordings from the in-lab and post-diary interviews, interaction logs from the diary study, and contextual notes participants provided alongside each interaction.

3.5.1 Participant feedback during in-lab session. We employed inductive thematic analysis [48], where transcriptions were first organized by the questions. The research team then added interpretative codes and notes through iterative discussions: first establishing a coding framework on a 20% sample, then applying and refining it across the remaining data. This process produced overarching themes capturing participants' expectations of ideal responses, preferences for customization techniques, and strategies for validating responses.

3.5.2 Annotation of diary study interactions. Participants shared links to their interactions, which included images and chats, and provided context. We annotated these data using deductive thematic analysis [33] and content analysis [43]. Predefined aspects included setting (indoor vs. outdoor; professional vs. leisure), familiarity with the environment/object/people, time sensitivity, and whether participants were alone or with others. We divided the annotation among three researchers. Before the post-interview, researchers pre-populated only those annotations that could be unambiguously inferred from the interaction itself (image and text) or from participant-provided context (e.g., clear indoor/outdoor cues; explicit references such as "my house" or "my ring doorbell" for annotating environment or item familiarity). Annotations that required information about participants' internal states or social context, such as whether they were in a hurry or whether others were present at the time of interaction, were intentionally left blank and were completed during the post-interviews by asking participants directly. We also coded the types of participant questions to the system, finding strong parallels with the categorization scheme of Brady et al. [18], where questions were instead answered by humans not AI. Last, we coded the customization techniques used. While most aligned with categories introduced in the lab, new techniques also emerged, which we categorized following the taxonomy in [84] (see Figure 6).

3.5.3 Participant feedback during post-diary interview. Before the post-interview, one researcher reviewed all participant-submitted

interactions for any potentially personally identifiable information (PII). The researcher also selected 5–6 diverse and discussion-provoking samples, which were later grouped by question type.

During the interview, we walked through all interactions one by one. For each interaction, the researcher informed the participant of any identified PII and explicitly asked whether they consented to share the interaction—including images, user prompts, and system responses—either in full, in redacted form, or not at all. We also refined and updated the diary annotations, adding contextual details that could not be inferred beforehand (e.g., whether the participant was in a hurry, whether others were present). Finally, the participant reflected on the selected interaction samples in terms of ideal responses, customization techniques, and comparing them across contexts and noting how their preferences shifted depending on the task or environment.

After the interview, we grouped the transcribed samples by participant and applied inductive thematic analysis using the same approach as in the in-lab study.

4 Results

From the in-lab session, we collected a total of 55 *Be My AI* interactions (five per participant) and 1,041 minutes of researcher-perspective video recorded via smartglasses (58:36 – 125:42, SD = 22:51), backed up by Zoom recordings via a laptop. In the diary study, we collected 363 interactions (22–48 per participant), typically accompanied by brief text message descriptions context and user goal. Participants often interwove daily reflections with these entries. We also recorded 1,232 minutes of post-study interviews in which participants discussed their experience in depth. In the Ask&Prompt dataset available at https://iamlabumd.github.io/ask-prompt/, we share 408 (418-10) interactions from the lab and diary study with our annotations. Ten interactions were omitted at participant's request, and 43 images/texts were redacted to remove personally identifiable information. Each interaction is manually annotated with details such as customization techniques, context, and user goal.

4.1 Characteristics of the Interactions

Efficiency is a critical factor in assistive systems, particularly for blind users who already face time disparities when navigating accessibility barriers in both physical [103] and digital [38] contexts. AI-powered VQA systems promise faster, near-instantaneous responses compared to volunteer sighted assistance, for more efficient, real-time access to visual information. Customization further enhances this potential by allowing users to tailor interactions to achieve their goals quickly [56]. However, estimating the time blind users spent per interaction in our study is challenging. Input methods vary (typing vs. dictation), each with different speeds [7, 47], and output speeds depend on individual screen reader settings [19]. We approximate time and effort using user and system word counts, contextualized by conversational turns, submitted images, goal achievement, and scenario characteristics.

Figure 3 shows a violin plot of interaction lengths for user inputs and system responses across both lab and diary sessions. The x-axis measures length in word counts. The width of each violin at a given count reflects the density of interactions—wider sections indicate more interactions at that length. The horizontal span shows the overall range of observed word counts. At a high level, responses are over ten times longer than user inputs. In the lab, user inputs have a median length of 20 words (range 0-64). Diary interactions show a lower median of 10 words but a much wider range, extending up to 258 words. This longer tail, likely reflects the nature of these real-world interactions, where participants had greater flexibility and autonomy in how and when they engage with the system—conditions that encouraged more detailed or context-rich input. The system is consistently more verbose than users in both settings. Responses have a median length of 242 words in the lab and 160 in the diary. Although generally shorter in the diary, the distribution shows a longer tail, where some diary interactions prompted more elaborated or multi-turn conversations.

4.1.1 User Input Length in Context. In the diary, 10% of interactions ($n = 38$) end with the system's initial response and no follow-up from users (input length is 0). Of these, only 3 involved a second image upload after a low-quality first image. In such cases, the response had 85 words on average ($SD = 37$). In nearly all instances, users reported that the initial response met their goal; the only exception was a low-quality image case, where the user chose not to continue.

In contrast, in 82% of diary interactions ($n = 302$), users followed up with inputs of 1-60 words. These interactions averaged 4 conversational turns ($1 - 21$, $SD = 2$) with 1 to 4 images submitted. In 60% of cases, users submitted only one image, suggesting that the first image was often sufficient for a useful response. System responses were longer in this group, averaging 207 words ($SD = 150$). Most interactions (83%) were considered successful by users.

Despite the overall success of most interactions, some were unsuccessful, typically falling into three categories: missing or inaccessible information, model limitations, and model errors. A common issue was that the needed information was absent from or unreadable in user images, often involving text on packaging. For example, P9 shared, *"It is very difficult to use this app with bottles. I could not get a clear picture. I even asked the AI how I could improve the picture and tried many different angles and distances, and I still wasn't able to get a clear picture."* Submitting additional images resolved 66% of such cases, though some questions remained unanswerable. Other failures arose when requests exceeded the model's capabilities. For instance, P6 asked about sofa dimensions to select a furniture cover—an inference the model is not capable of making. In fact, we noticed a persistent trend where the system avoided guessing measurements and distances in space (*e.g.*, when P7 asked, *"If you were to guess, how deep is the snow?"*). Several participants sought advice on medications, which the system is not qualified to provide. In these cases, the model appropriately directed users to professionals. Finally, some interactions failed due to model errors. Users often recognized inaccuracies when responses contradicted prior knowledge or did not hold up in practice. For instance, when P10 asked for the expiration date on a bag of salad, the system responded with "February 2024." P10 noted, *"I knew it couldn't be that old."* In contrast, a few participants accepted incorrect outputs—typically in text-reading tasks involving complex layouts, such as microwave control panels or washing machine knobs—where verification was difficult and users defaulted to trust.

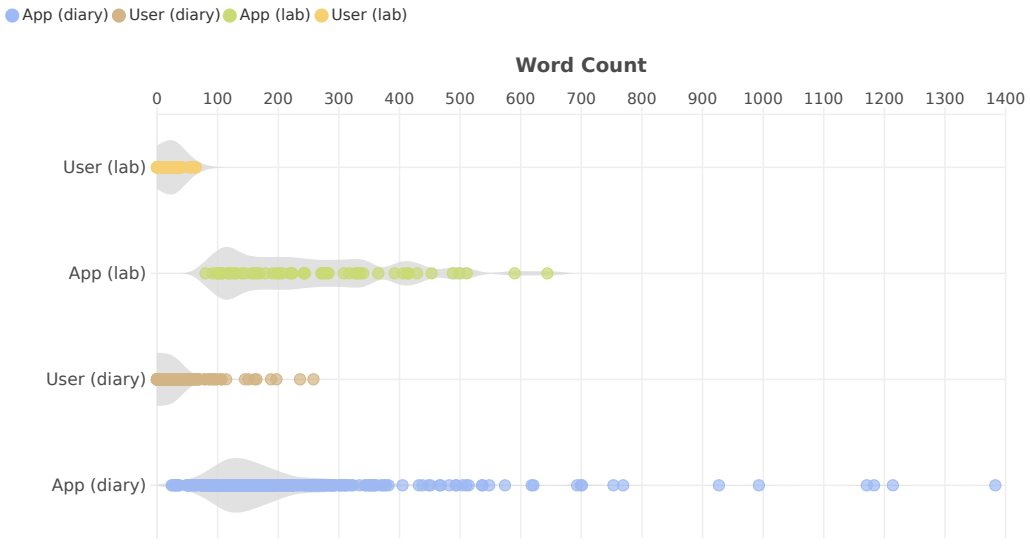


Figure 3: Lengths of user inputs and system responses (measured in word count) across in-lab session and diary session interactions.

We also observe a long tail of interactions ($n = 27$, 7%) with substantially longer user inputs (60- 258 words). These typically involve more conversational turns, with a median of 6 turns ($2 - 17$, $IQR = 3 - 12$), reflecting deeper dialogue. Extended exchanges arose when users made multiple attempts to capture usable images, required multiple clarifications, asked curiosity-driven follow-up questions, or requested updates on dynamic scenes. We further explore this iterative process in the next section. Not surprisingly, these longer interactions included more images (1-11, median=3) and longer overall system responses (36-1383, median = 309). Despite their complexity, 78% were considered successful; the unsuccessful cases involved complex tasks combined with model errors, leading users to give up after extended attempts.

4.1.2 Patterns in Conversational Turns. To examine structural patterns in how participants progressed through their interactions, we applied k-means clustering ($k = 3$) using two features: the number of conversational rounds and whether participants reported successfully completing the task. We selected k based on interpretability. This analysis on the diary data, surfaced three distinct clusters that reflect common patterns of turn taking and reported outcomes.

Cluster 1: Short-round interactions reported as successful (n = 286). The majority of the interactions involved only a small number of conversational rounds (median = 2 rounds, range = 1–7) in which participants indicated that they obtained the information they needed to complete their task with only light prompting. All participants had at least one interaction that fell into this cluster. An example of this appears in tasks related to identification and visual exploration. For instance, P8 captured a clear photo of the view outside an airplane window. The system correctly described the scenery, and after she responded with appreciation, it replied with *"I'm glad I could help…"* (a total of 2 rounds). Reflecting on this interaction, she explained: *"I got the response that I was looking for. I wanted to know what was going on outside of the window seat."*

Cluster 2: Short-to-medium-round interactions reported as unsuccessful (n = 53). These interactions consisted of a modest number of rounds (median = 3, range = 1–9) but were reported as not meeting participants' needs. Participants rarely used any prompting and often stopped early, which may indicate that they encountered issues they felt unlikely to resolve within that interaction. Interactions in this cluster came from only about half of the participants (P1, P5, P6, P8, P9). We typically observed this type of interaction when the camera frame did not capture the information needed to interpret the scene. For example, P6 made 4 rounds with Be My AI while asking for directions on using a razor. Despite taking three images, she was unable to include the part of the box that containing the instructions and ultimately discontinued the interaction.

Cluster 3: Long, effortful dialogues with mixed outcomes (n = 24). These extended interactions (median = 10.5 rounds, range = 8–21) showed the highest use of customization techniques (median = 2 techniques, range = 0–8). Participants described these cases as requiring substantial troubleshooting or refinement of outputs, and outcomes varied. Interactions of this type came from six participants (P9, P6, P2, P11, P3, P10). These appeared to be moments where participants relied more heavily on prompting techniques to guide the system toward responses they considered more useful. They typically involved participants feeling encouraged to continue when the system's reassuring or seemingly accurate responses signaled progress. For example, P2 used decomposition and grounding while searching for ranch dressing in her refrigerator, continuing for 10 rounds until she located it. As she noted, *"I knew that the app was correct."*

4.1.3 Scenario Characteristics as Context. Based on participants' text descriptions of their contexts and goals, we annotated the scenarios in their interactions and verified these annotations in post-interviews. The bar graph in Figure 5 shows the distribution of all interactions across different scenario characteristics (y-axis).

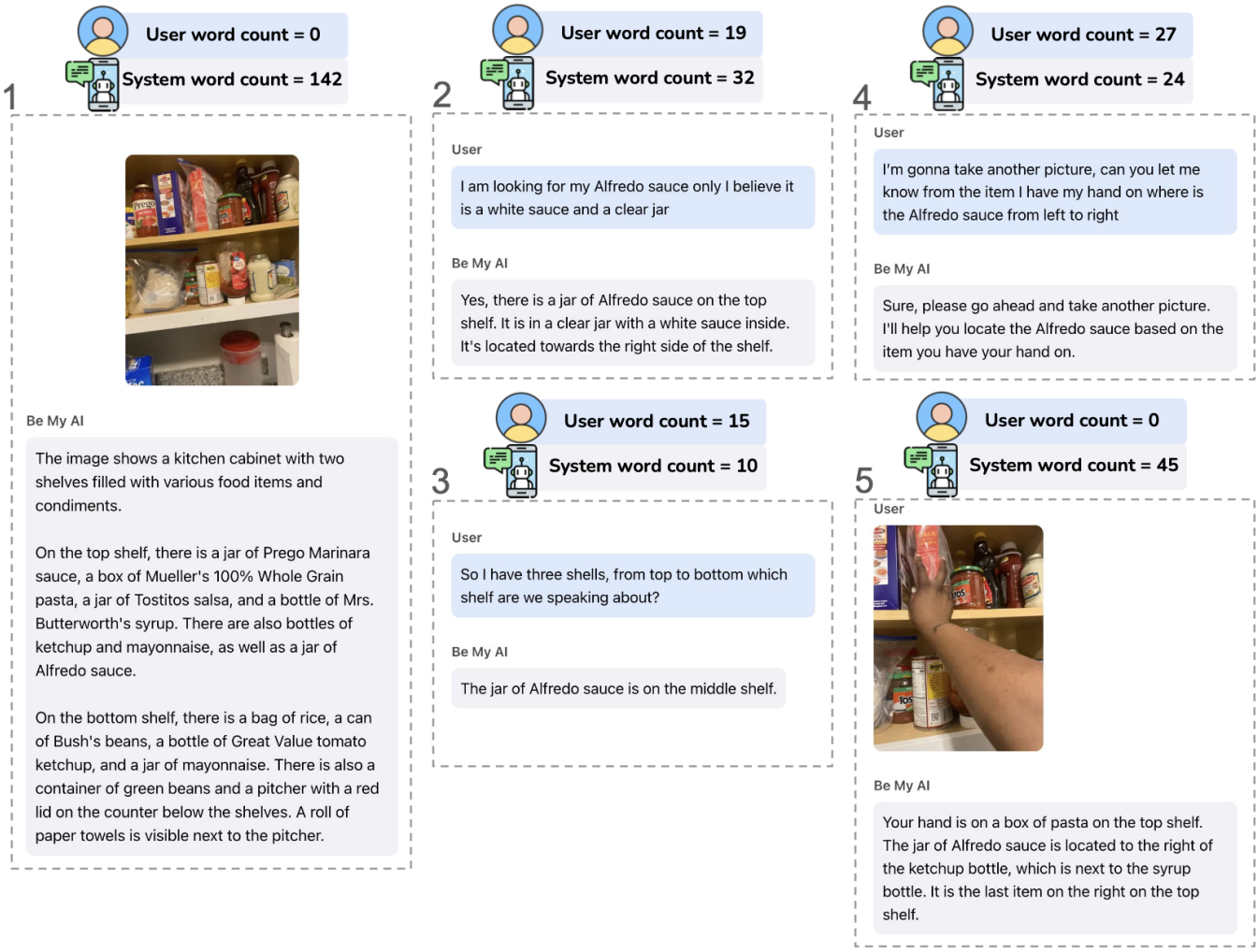


Figure 4: Example interaction consisting of five conversational turns with word count for each turn.

Most cases reflect a clear presence (red) or absence (blue) of the characteristic, with relatively few uncertain cases (gray). Bar lengths indicate the percentage of interactions in each scenario. Most logged interactions occurred during leisure activities (93.4%) and when participants reported being not in a hurry (84.8%), often in familiar environments (87.9%) or indoors (80.7%). Social context was also associated with when interactions took place: most occurred when participants were alone (67.2%) with fewer interactions happening around familiar people (17.1%), strangers (11.0%), or mixed groups (4.4%). Similarly, most interactions involved familiar items (72.5%). Overall, participants used the conversational VQA system mainly in everyday, comfortable routines rather than novel or high-stress situations. It remains unclear how lengthy or erroneous system responses influence these choices.

The color-intensity segments of each bar in Figure 5 show the proportion of interactions that involved user text input and customization under each scenario. At a high level, we observe that participants tried customization techniques more often in familiar environments (55%) than in unfamiliar ones (34%). A similar pattern emerged with items: 54% of interactions with familiar items involve customization versus 47% with unfamiliar items. Social context also mattered but in a different way: participants customized most when surrounded by familiar people (60%), slightly less with strangers (56%), and least when alone (49%). Physical setting had little effect, with similar rates indoors (52%) and outdoors (54%). Customization occurred in 46% of hurried interactions, compared to 54% when participants were not in a hurry. In professional contexts, participants customized far more (74%) than in leisure settings (52%), an unexpected finding given that casual environments might seem more conducive to exploration.

4.2 Participants' Tactics for Tailoring Responses

In this section, we examine how blind users interact with the VQA system, focusing on how they attempt to control and tailor responses to their needs. Our analysis of the interactions–both in the lab and participants' diaries–along with their reflections, reveals

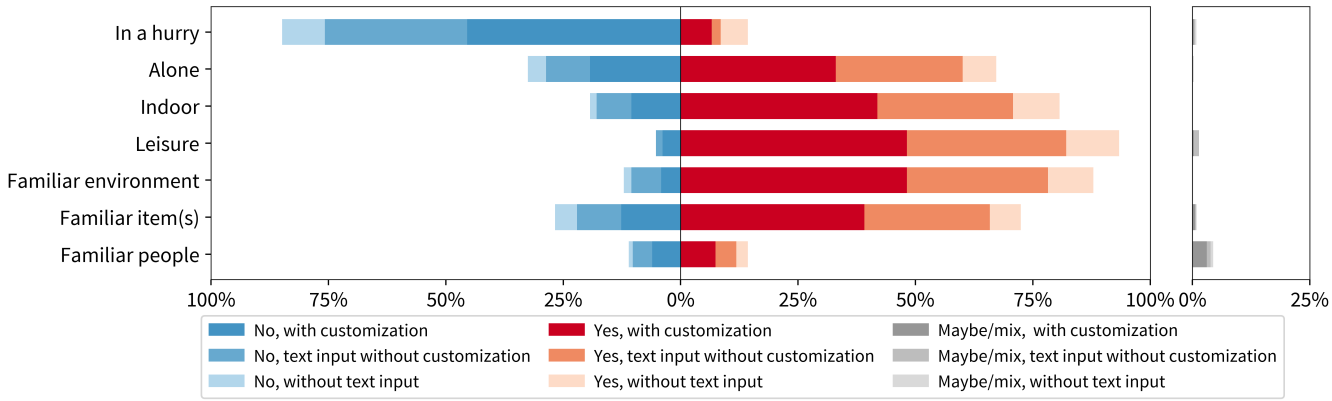


Figure 5: Distribution of interactions across scenarios, with colors indicating presence, absence, or uncertainty and intensity marking customization. Proportionally, interactions involved more customization in familiar environments and with familiar items, slightly more around others than alone, similar indoors and outdoors, and less when in hurry.

several patterns. By default, the system provides an image description with each input, but participants often find these descriptions lengthy, broad, or insufficient for completing a specific tasks. The most common strategy is to ask direct questions without tailoring. Beyond this, participants carry over prompting techniques from the lab and experiment with new ones, as shown in Figure 6. Through these approaches, they sought to control verbosity, direct attention, and better achieve their goals.

4.2.1 No Prompting Technique. We observe that the most prevalent way blind users interact with the system is by simply asking questions without explicitly attempting to further tailor the answers. This is the case even in the lab, where participants were instructed to prompt and customize. In some cases, participants mentioned that customization techniques did not always feel intuitive and that there was a learning curve. As P3 explained: *"I don't think it will come natural. I mean, I even forgot about that aspect of it because when I'm trying to get information, usually the way to go is to ask for it. But I'm not close-minded to that personalized technique. .. I would have to get more familiar with it to use it."*

On average, 77.27% of each participant's in-lab interactions included at least one question (50%–100%, $SD = 20.78\%$). In their diaries, where they chose use-case scenarios of their own, participants still asked a high proportion of questions, with at least one question appearing in 74.8% of their interactions on average ($0\% - 100\%$, $SD = 29.81\%$). Both in the lab and in their diaries, we observe that with their questions, all ($n = 11$) participants shape the responses to be more task oriented. For instance, in the first in-lab scenario of locating the door, participants are instructed to provide binary feedback related to the helpfulness of the description in the response. Yet, P2 asked *"Where is the exit sign?"* after the first turn of interaction, aiming to narrow the response to help determine the general direction of the exit door.

Indeed, in its current form, *Be My AI* defaults in image description without a context for a task. Users are not allowed to submit a question alongside the first image, but through subsequent interactions. More than half ($n = 6$) of the participants found this limitation to be critical as it often forced them to sit through lengthy descriptions before getting to the information they actually needed. Reflecting on this after trying to locate her keys in an unfamiliar room, P6 shared: *"when I took a picture, it gave me a bunch of information that I did not want. So I have a question [in mind] before I take a picture."* Even when the goal was to obtain a description, in many of their interactions (40.38%) participants follow up the default initial description with questions for more. For instance, P3 asked *"What color are the pillows on the couch?"* for detailed descriptions after knowing there were several decorative pillows on the sofa from *Be My AI*'s default description. Participants also ask *Be My AI* to read specific text in 20.33% of their interactions, often supplementing with additional images. For example, P10 asked *"What is the expiration date"* of a yogurt with several images to let *Be My AI* find and read the date.

As shown in Fig 7, 9.62% of the interactions in the diaries involve questions for object localization and orientation, while 8.79% of interactions involve questions for object identification. Interestingly, with similar breakdown by Brady et al. Brady et al. [18], our distribution notably differs from that of VizWiz. The large drop of identification questions from 44% in VizWiz to 8.79% in our study could be explained by *Be My AI*'s default image description right after the image, preempting the need for identification questions. We also suspect that *Be My AI*'s support for real-time and multi-turn question answering explains the lower number of unanswerable questions from 7% to 5.49%, with an additional 10.44% of unanswerable cases in the diaries being resolved through the submission of additional photos. More so, we observe the emergence of a new question category pertaining to localization despite *Be My AI* terms of service indicating that it is "not designed or intended to be used as a mobility aid" [31].

Unsurprisingly, even though the purpose of the study was to investigate customization, asking questions was deemed a more natural and intuitive way for the blind participants in our study to interact with this VQA system. P7 preferred asking question to framing them with prompting techniques because asking question is *"just more natural ... like having a conversation."* P3, who did

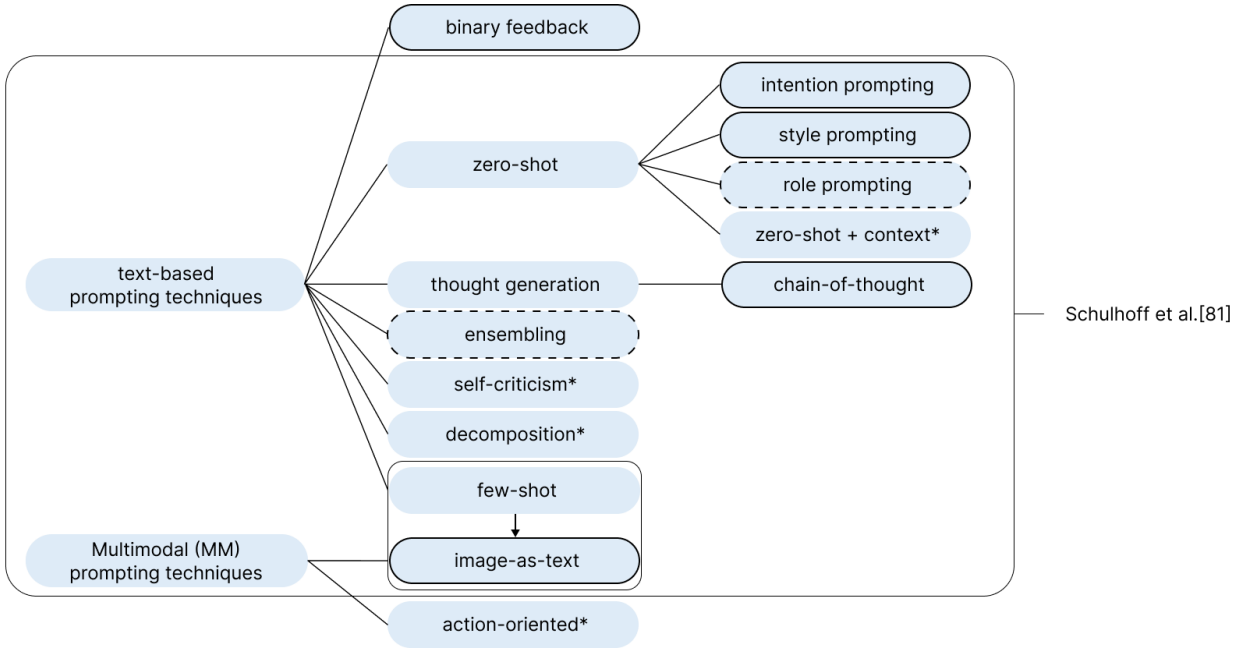


Figure 6: All the customization techniques observed in Ask&Prompt dataset following the hierarchy from [84]. The techniques with solid border are introduced in the lab. Techniques with asterisk come up by participants in the diaries. Techniques with dash border are customization techniques that participants' explored strategies resemble. Due to the limitation of *Be My AI*, we use image-as-text prompting instead of few-shot prompting to provide information.

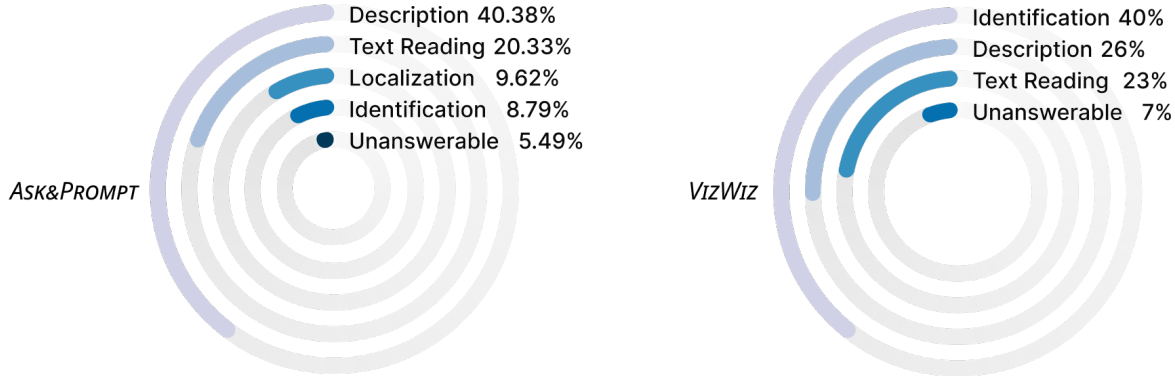


Figure 7: Distribution of question categories in 363 multi-turn interactions from participants in our diary study session and 495 single-turn interactions randomly sampled in VizWiz [18].

not use any technique but asked questions in 56.75% of her diaries, concluded her way of interacting similar to *"googling something"* because she *"always ask a question"* to get the information. Although she acknowledged that, compare to asking questions, prompting techniques could, in theory, make responses more concise or relevant, she did not think they would come naturally. Yet, she still stayed open to the idea of using prompting techniques in the future and said, *"I would have to get more familiar with it (prompting techniques) to use it... I hasn't transitioned [my way of interacting] in my mind yet with artificial intelligence."*

4.2.2 Binary feedback. This is the first customization technique introduced in the lab situated within an exit door navigation task, where participants could provide one-word feedback on individual responses during the interaction. [2]

Binary feedback is a common interaction pattern across many interfaces and does not necessarily imply system adaptation. We introduced it as an accessible baseline to anchor discussions about

[2] *Be My AI* only supports feedback through a "thumbs-up" or "thumbs-down" feedback button at the end of the chat but not after each individual interaction response unlike other commercial interfaces, e.g., ChatGPT, and thus this feedback mechanism does not directly impact the user's ability to achieve their immediate goals.

feedback mechanisms and to contrast it with more expressive strategies. Its simplicity helped frame participants' expectations about what "feedback" could mean, even within a system without memory.

Almost all (n = 10) participants practiced this technique after our lab introduction. Surprisingly, although many (n = 8) of them gave *"helpful/great"* feedback, some of them did not find the description they received sufficiently helpful. For example, P3 gave a *"helpful"* feedback, even though she did not manage to locate the door using the app's responses alone. One possible reason for this is that participants may have wanted to maintain a positive attitude toward the technology they were testing [96]. As being the easiest prompting technique, however, more than half (n = 6) of the participants did not prefer using it immediately in the lab. One key reason is its limitation of lack of specificity. For example, P5 said, *"if I just say helpful, okay, helpful on what? I don't express, it's like nothing."*

In their diaries, we observe that about half (n = 6) of the participants made limited use of this technique, though distinct usage patterns still emerged. P7 indicated that binary feedback is more natural to be provided at the end, which is similar to what *Be My AI* has implemented by the "thumbs-up" or "thumbs-down" feedback button at the end of the chat. With greater flexibility during the diary study, half (n = 5) of the participants expressed binary feedback at the end of the interactions.

Another observed pattern is that, in their diaries, participants naturally paired binary feedback with a short rational, addressing the lack of specificity they had noticed in one-word binary feedback during the lab study. This feedback style emerged organically through casual conversation with the system rather than as a deliberate strategy. For example, P8 spontaneously shared her excitement with *Be My AI* after discovering details outside her airplane window: *""Wow, this is my first time scanning a picture and realizing that there's a plane parked next to the plane that I'm currently sitting in. I appreciate all the details."* Similarly, as shown in Figure 8, P11 gave a *"very helpful"* feedback along with the detailed reason. He further explained in the post interview that this kind of feedback is not always intentional and strategically planned: *"It (the system) would just evoke a natural response."* He further explained that a more human-like voice, such as the one used in PiccyBot [80], could further encourage this behavior: *"It's literally like you're talking to a person. So it almost invites you to respond, even if you're annoyed with it, or if it did a really good job, saying wow, that's awesome, that's so helpful."*

4.2.3 Zero-shot: style prompting. This is the second customization technique that participants are introduced into within the exit door navigation task. About half (n = 6) of the participants tried it in the lab following our prompting examples using *"less details"* or *"more details"* to **control the verbosity** of responses. Some (n = 3) went further by adding additional questions or instructions guiding the system to filter the information for them. For example, P8 prompted with, *"Less details,.. just tell me what is very close to me"*.

In their diaries, we observe a similar usage pattern. More than half of the participants (n = 6) used style prompting paired with specific instructions or other techniques to control the verbosity of the responses. They primarily used it to receive shorter, more focused answers. For example, P10 repeatedly framed prompts as *"Tell me only about ... and nothing else around it"* using this pattern each time to filter for relevant information. Indeed some participants (n = 4) explicitly reported that they would use style prompting as a filter to *"eliminate the unnecessary stuff"*. P2 described this approach as intuitive, likening it to natural conversation: *"that's just like having a conversation with somebody, and they're going on and on and on, they're long-winded."* P3, who had prior experience with *Be My AI* for months, immediately found style prompting useful in the lab because it finally addressed the issue that *Be My AI "gives way too much information sometimes."* However, she did not use this technique in her diaries. We suspect that the reason is still her unfamiliarity to techniques mentioned in 4.2.1.

4.2.4 Zero-shot: intention prompting. We introduced intention prompting during the nutrition-fact identification task in the lab, clarifying that it can reduce repetition and streamline interactions, particularly when users have multiple queries tied to the same underlying goal (e.g., dietary assessment, cooking progress, etc.). Although most (n = 9) participants practiced it in the lab, some initially struggled because they were accustomed to phrasing their intentions as direct factual questions. For example, P6 began with, *"It's important for me to eat lower carb and high protein, how many grams of protein is this?"* After further guidance, she refined the prompt to convey her broader goal and invite an open-ended response: *"I follow a low carb diet. Is this a good food for me?"* In their diaries, we observe that this habit of interpreting intentions as concrete questions persisted. For example, P8 framed her intention of deciding whether to wear high heels to a concert as a concrete question asking whether the outside was wet. Interestingly, she independently noticed this during the post interview and proactively improved her prompting style, explaining: *"when we are more intentional in what we're asking, then we're going to get more specific responses."*

Almost all (n = 8) participants who tried this technique in the lab said they would use it again and even envisioned future use-case scenarios. For example, P5 imagined applying it while grocery shopping: *"I have more details and some advice... that can make me start to make my grocery by myself."* Similarly, P3 pictured using it to choose a lipstick color independently. Indeed, in their diaries, one of the main uses of intention prompting was for requesting suggestions. For instance, P6 shared a photo of her kitchen and prompted directly: *"I want to remodel my kitchen, I need suggestions"*.

In the post interview, many participants (n = 8) mentioned their preference towards intention prompting. For example, P7 explained that this technique helped *Be My AI "understand where I'm trying to go, as opposed to me having to direct it [pause] exactly the specific point."* However, initial reactions varied. In the lab, P10 and P11 both expressed reluctance to use intention prompting, expressing confidence in their own questioning skills and limited trust in the system's ability to infer intent. P10, focused on reducing sodium intake, said: *"I don't think I would tell the app that (my intention), I would just ask how much sodium is in here... cause I don't think it's necessary to get the answer that I want."* Interestingly, both later adopted intention prompting repeatedly in their diaries and named it a favorite technique in the post interview. P10 described it as

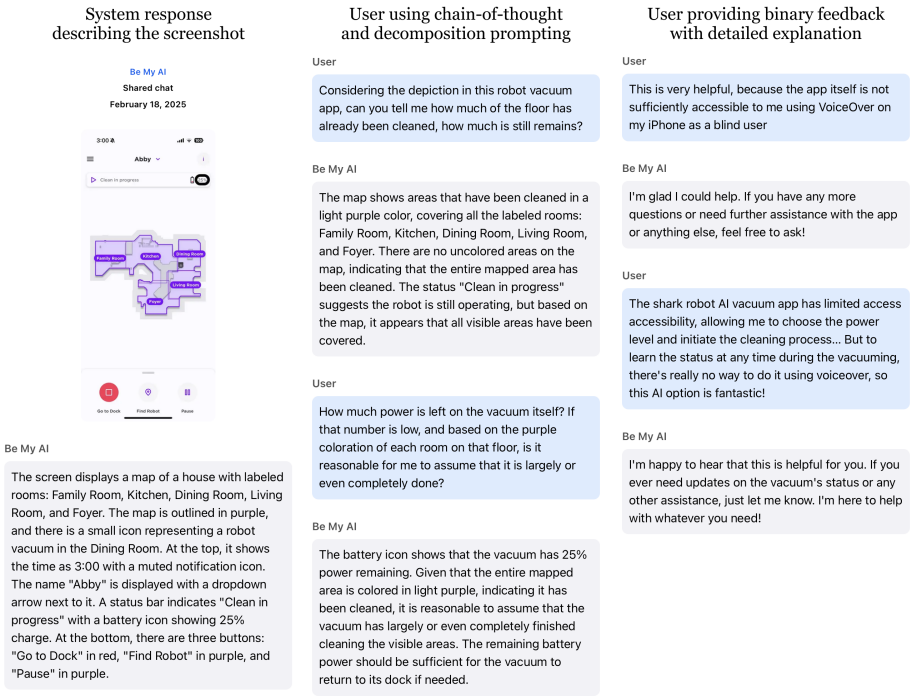


Figure 8: P11 using (a)) binary feedback along with detailed explanation in the style of chatting, (b) chain-of-thought prompting during the diary session, and (c) explicitly mentioning vision level.

feeling *"natural"* to use while P11 appreciated its efficiency: *"because you can certainly narrow the lane dramatically [by expressing the intention] ... and you'll get to the finish line much more quickly."* P11 also discovered that *Be My AI* could carry his stated intent across follow-up images, calling this *"a big time saver which I hadn't appreciated."*

4.2.5 Chain-of-thought (CoT) and decomposition prompting. CoT prompting is introduced to participants in the lab within the item localization task by encouraging them to include reasoning behind their queries. We observe that this technique appears to be the most challenging for participants to grasp. In the end, only about half of the participants ($n = 5$) attempt to use it in their prompts. For example, P8 incorporated reasoning by prompting *"I want to know if there are any cups with liquids near the Lipton soup because I don't want to spill anything."* Even after researcher's explanation and examples, some participants ($n = 4$), still defaulted to either style prompting, intention prompting, or simply asking questions. Use of CoT remained limited in their diaries as well. One notable exception was P11, who repeatedly applied CoT when interpreting screenshots. As shown in Fig 4, he guided *Be My AI* to integrate multiple visual cues (e.g., color coding, battery level, room labels) to infer a vacuum's cleaning status. We speculate that his familiarity with CoT may be attributed to his extensive experience interacting with *Be My AI* and other GenAI systems.

Although we introduced CoT in the lab, three participants showed at least one instance of interpreting it as decomposition prompting, breaking a task into subtasks, or a series of questions to guide the system. For example, P9 decomposed the item localization task into two parts by prompting *"I'd like to know where the soup is and what objects are around it."* This technique is even more common in their diaries, where more complex real-world scenarios occurred. Many ($n = 9$) participants decomposed tasks to get more precise, step-by-step responses. For instance, P2 decomposed her task of locating a food box on a shelf by asking: *"From top to bottom which shelf is it located on, and from left to right exactly where is it located? Is it visible or behind something?"* P9 also applied this technique in cooking scenarios. As shown in Figure 9, she breaks down the assessment of dough readiness into several descriptive questions to gather detailed information.

Many participants ($n = 8$) expressed a clear preference towards decomposition prompting during the post interview. P7, who used decomposition prompting in the majority of his diaries, emphasized that this technique was particularly useful when dealing with deviations, instances when responses began to drift from the goal he had in mind. He explained that breaking the task into smaller steps let him pull the conversation back on track. In addition to this, participants valued decomposition prompting as a way to obtain more complete and accurate information. P9 noted that although it can be time-consuming, breaking down the task is *"the only way to make sure that it (the response) was 100% correct."* Similarly, P11 appreciated that decomposition prompting enabled him to get granular details without relying on the system's own filtering, especially for image description. He adopted this technique after the AI once omitted elements it judged *"inappropriate,"* rather than describing everything visible *"as objectively and neutrally as possible."* As many

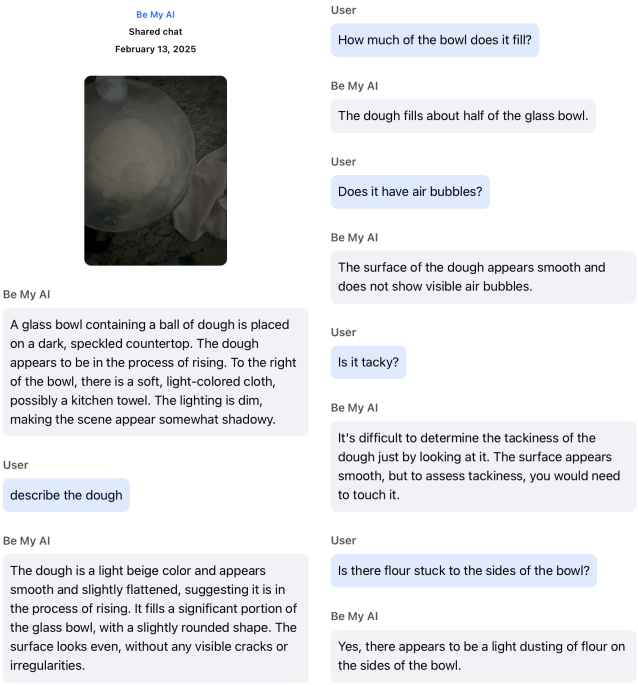


Figure 9: P9 using decomposition prompting in the diary.

participants independently came up with this technique in their diaries, they later reflected on its naturalness in the post interview. For example, P9 found her interactions with *Be My AI* similar to those she had with sighted volunteers unfamiliar with the context, which led her to apply the same step-by-step prompting she typically used with people to her interactions with the system.

4.2.6 Image-as-text prompting. This is the last customization technique that participants are exposed to in the lab, situated in the personal item localization task. Similar to teachable object recognizers [53, 95], participants envisioned that it will be useful to locate personal items or distinguish between similar objects. For instance, P10 imagined to identify her cane at a large conference and said, *"If you ever been to a conference where there are 5,000 blind people, and you want to know which cane is yours. In that case, I will put something on my cane, like some kind of keychain or charm, so that if they do get mixed up I know which one is mine. So that's a question that I could then ask."*

In their diaries, we observe similar use patterns to distinguish objects. For instance, P11 used this technique to identify a sanitizer bottle next to a similar one by referencing a known visual detail: *"The sanitizer has the blue cap."* Participants also used this technique to shape the focus of the interaction. For example, P8 prompted, *"I am the woman in the blue and white striped dress holding a white cane,"* to narrow down the focus to herself. She then added more specific details, e.g., *"coconut booties"* and *"the silver Tiffany stud earrings in both my ears"*, to check each item she was wearing, ensuring that she looked good in the photo.

Almost all participants ($n = 10$) expressed a preference for this technique and anticipated using it in the future when asked in the lab. Yet, this technique appeared in a very limited number of diaries (6.61%). We theorize this contrast is explained by the need for one to be able to recall detailed visual characteristics of objects or scenes for them to be described in detail to *Be My AI*. For instance, participants often looked for or located items they could typically identify by name or recall a few key properties. For example, when P2 asked *Be My AI* to look for a Pop-Tart on the shelf, she was uncertain, *"maybe strawberry flavor."* In the post-interview, she expressed a willingness to improve her prompt using image-as-text prompting, yet still struggled with not knowing the appearance and having nothing specific to describe. This difficulty even arise with personal items that participants use everyday. For instance, P3, the only participant who did not practice this technique in the lab, felt uncertain about the appearance of her keychain and found it hard to describe. She said, *"I don't think I would naturally do that. Oftentimes, maybe I don't know the details of it."* We believe that few-shot prompting, the original technique we wanted to introduce and not supported by *Be My AI* could help overcome unfamiliarity with object and scene appearance even for familiar objects that may be the key reason for the limited use of this technique.

4.2.7 Customization techniques beyond the ones introduced in the lab. Beyond the customization techniques introduced in the lab, we observed that participants spontaneously combined zero-shot prompting with rich contextual information as a form of grounding. Regarding the reason of providing the rich context, P2 said, *"The more feedback or content you give it, the more feedback and content that it will give you."* She concluded it as a way to prompt the system to generate more accurate and relevant responses. In many cases, the contextual information participants provide not only supports

the current task but also implicitly signals personal preferences or patterns that could improve alignment over time.

In the diaries, participants also encountered situations where *Be My AI* provided responses that were clearly incorrect. In these cases, participants naturally employed self-criticism prompting to address sycophantic behavior by questioning and correcting the original response[84]. For instance, after *Be My AI* responded with a price of $699.90 for a dog bed in a supermarket, P7 questioned the system by asking, *"Does the price actually say* $69.99*?"* In some cases, participants followed up by asking the system to generate the response again after issuing a correction. We observed that *Be My AI* provided correct information after participants used self-criticism prompting technique. However, it remains unclear whether the system truly revisited the image to verify the correction. Participants still face the risk that the system simply overagree with their correction rather than independently validating it.

Participants also explored prompting techniques that resembled widely used customization techniques. One such strategy involved explicitly stating their blind identity in the prompt. In fact, neither researchers nor participants know whether Be My AI uses a custom system prompt or dataset; Surprisingly, we observed cases where explicitly revealing blindness changed system behavior. For example, P9 prompted the system to shift the reference point from the image framing to surrounding objects by saying, *"I need to know what objects are around it because I'm blind and can't see what is framed in the photo."* In response, the system provided more comprehensive descriptions of nearby items and their relation to the object of interest. Although participants did not explicitly assign a role to the system, their use aligns with role prompting [84]. Thus, we suggest that role prompting can be a possible strategy for blind users to better align the system's responses with users' needs and expectations.

In addition to strategies resembling commonly used customization techniques, participants came up with a new prompting technique—including their action in the prompt. Typically, they prompted the system with the action they were about to take in the next image. For example, P2 prompted the system before uploading an image with her hand in by *"I'm gonna take a picture of me touching a bottle, from that picture. Please let me know if I need to go left right up or down."* This action-oriented customization strategy illustrates how blind participants use prompts to shift the system's focus in alignment with the evolving steps of a real-world task. This also shows the potential of VQA systems to support ongoing, real-time activities through conversational interaction.

4.3 Other Opportunities for User Control and Customization

In the previous section, we saw the customization techniques that participants used; some of them they were exposed to and some of them they were able to come up with. In this section, we identify other opportunities for customization that while might have not appeared in participants' interactions they could be useful based on participants feedback on the ideal response that they would expect. We discuss these in the context of the tasks that the participants were trying to perform since ideal responses among tasks shared common themes.

4.3.1 Localization tasks call for verbosity control and personalization of spatial points of reference and safety cues. In the lab, participants consistently valued detailed descriptions when navigating unfamiliar environments, as these helped them build a clear mental map of their surroundings. As P4 put it when locating the exit door in the lab: *"I like more information to know what's around me."* However, in their diary interactions, where localization tasks were often performed under time pressure, preferences shifted. In these hurried contexts, participants most appreciated concise, task-focused responses and expressed frustration when the system added information irrelevant to the immediate goal. For instance, P1, when searching for her TV remote at home, explained: *"If it was able to just say… we don't see the remote, that would have been helpful so I could take the pictures faster without having to cipher through everything."* This contrast strengthens our previous observation (Section 4.2.3) on the need for **verbosity control**, where users can toggle between detailed or concise output depending on context.

A second theme concerned **points of reference** in spatial descriptions which was found both in the lab and diary interactions. Some participants preferred object locations described relative to themselves or relative to their hand when pointing, aligning with prior findings that blind users commonly include their hand, either touching or placing it near the target, as a reference during localization tasks [60]. For example P2 wanted responses like *"to the left of your hand"* often paired with distance cues such as *"... from where you was just positioned, go to maybe the left, or… take five steps"*. Others preferred locations described relative to familiar landmarks. For example, P11 said, *"If you knew where something was, like the laptop… you could use it as a fixed point of reference"*. These preferences contrasted with *Be My AI*'s defaults that were relative to the image frame (e.g., *"top left of the image"*), which participants found less accessible and thus less actionable.

Participants often expressed a desire for **safety awareness** beyond the target object with information about nearby or potentially hazardous items to support safe navigation. As P8 explained, *"I hate the fact that… sometimes we just reach out into space… people reach and they knock you in the head, they've spilled stuff on you."* This emphasis on environmental caution echoes prior findings [37], who identified the avoidance of dangerous or unwanted objects as a frequent use case for blind participants using AI scene description tools.

4.3.2 Description tasks call for detail-level control and personalization on attributes, order, and time-lapse. These tasks involved requests for information about objects, scenes, screenshots, or people. In these cases, participants typically valued factual and relevant information. We also saw opportunities for customization in defining **attributes of interest**. For example, P11 wanted fine-grained color distinctions when asking about clothing: *"It's particularly helpful if it tells me the shade of the color … there must be dozens of shades of blue that can be discerned, and they can be quite different."* By contrast, P10 focused on broader color categories (e.g., black, red). Prior work suggests detail preferences may relate to vision history [88]: P11 lost sight later in life, while P10 was blind from birth. Such differences highlight potential for system-level customization through **memory of user preferences** [74], enabling responses tailored across sessions.

Echoing prior findings [71], participants also expressed preferences for the **order of information**. For example, P6 preferred cabinet descriptions to begin with style and color before number and layout. Beyond verbosity control, this points to the need for tailoring what kinds of details appear at each level. When no fixed preferences existed, participants favored a **hierarchical response style**: starting with a high-level overview, then offering ways to drill down. Dense outputs could be overwhelming (P7: *"If it's just a lot of information, I can't process at all"*), but participants appreciated when the system inferred intention and offered follow-ups (P11: *"I would have liked the overall perspective first and then have it go on its own to a little more detail. Without me having to."*).

Participants also valued **time-lapse estimation** with descriptions linked to actionable next steps. P7, after photographing sausages, imagined the system saying: *"Maybe it still looks pretty pink—stir a little more, maybe about five minutes."* This mirrors preferences in localization tasks, where relative distances (e.g., *"a few steps to the left"*) helped guide actions. However, visual temporal understanding such as temporal order understanding and time-lapse estimation is limited in state-of-the-art VLMs [51].

4.3.3 Identification tasks call for automatic verbosity control and personalization of proactive advice and camera guidance. Identification tasks centered on naming or recognizing objects, products, or items, often to decide how to use them or to distinguish between similar options. In these tasks, participants generally prioritized conciseness, favoring short, direct answers over lengthy descriptions. Unlike description tasks, identification in diary interactions often involved unfamiliar items, which shaped participants' expectations of what counted as an ideal response.

In many cases, participants prioritized **binary confirmation** over lengthy output. Rather than full-scene descriptions, they wanted simple yes/no answers, such as whether a bag contained broccoli (P3), whether snow was on the grass (P5), or whether a park bench was empty (P5). Building on prior work [88] that found the amount of time the user has influences the amount of content they want, some participants suggested a **"hurry mode"** toggle, providing only the bare minimum in time-sensitive contexts. As P11 explained about the interaction when he was getting ready to go out: *"If you had a hurry button and I was in a hurry, that probably would be OK. If it just said no stain, it's good to go."*

At the same time, some participants welcomed **proactive advisory suggestions**. Clothing identification often invited proactive fashion advice (P10: *"It could say, Well, that does match. But wearing a red shirt would look better."*). Others, like P6, imagined instructions for use following identification, such as how to operate a product once recognized.

Another common need was **proactive image framing**. In most cases where participants struggled with image capture, the underlying issue was improper focus or cropping of the object. Several participants explicitly asked how to take a better picture, yet the system's camera guidance remained generic—providing suggestions such as *"ensure the room is well-lit," "hold the camera steady,"* or *"adjust the angle to capture the full screen."* which often did not address the primary problem. Participants expressed a clear need for more actionable, object-aware feedback. As P7 explained, *"It would be useful if there was some kind of focus indicator… with rounded bottles or cans it tends to be harder to capture the image in a way that is interpretable for the AI… Just like when I'm taking a regular picture with the iPhone and it says 'centered,' 'tilt left,' 'tilt right.' Tools like that would help take better pictures to get better responses."*

Finally, several participants emphasized the need for **automatic verbosity control**, where the system could remember preferences and adapt detail level across contexts. As P7 noted, *"In an ideal AI situation, where it does learn your preferences... it would still probably need some adjustment in different scenarios, but for the most part it could possibly get."*

4.3.4 Text reading tasks call for personalization of content priorities and hierarchical summaries. These tasks involved reading information from product labels, menus, screenshots, and other written content. In the lab, participants consistently stated that the system provided too much information when reading entire ingredient lists on product packages. Almost all aparticipants expressed preference towards specifying question before, or along with the image input, allowing them to clarify what they wanted in advance. As P10 explained: *"Even before I can take a picture, I could ask the question, how much protein is in this product? So that I don't have to listen to all this other stuff that I don't really care about."* Prompting techniques like **intention prompting**, introduced in the lab, could similarly help streamline outputs.

In diary interactions, additional themes surfaced. Similar to description tasks, participants expressed preference for a **hierarchical response style**, starting with a high-level overview or summary and then offering the option to drill down for details. This mirrors screen reader conventions, where users skim at a fisheye view before zooming in. They also believed this approach could reduce errors from the system missing content. For example, P9, who tried to read a large restaurant menu, found that sections were repeatedly omitted. She suggested: *"It could have listed the sections, appetizers, entrees, desserts, and then asked what I wanted to know about. That minimizes the risk of missing sections."* She added that a high-level overview *"would have been nice to know what all of them were, instead of having to ask several times."* Related prompting techniques like **decomposition**, breaking the task into subtasks (e.g., reading menus section by section), could also support this goal.

Participants further emphasized the importance of **comprehensive product knowledge**, especially when reading text on packages. As P5 put it: *"It would have been good to put everything there [in the database], so that when you have a product, maybe you take a picture, you can ask from that picture many information without taking several pictures."*

4.4 Participants Trust and Verification Practices

4.4.1 Localization. **Trust** in localization was generally lower than in other tasks. Participants were cautious about relying on the system for navigation and distance judgments, especially under time pressure. P2 did not trust it for *"moving from one area to another,"* and while they were willing to use it alone in daytime or in unfamiliar areas, they avoided doing so at night. Safety concerns further limited trust; for instance, P1 preferred human help when

cones blocked her path and P3 explained she would not use it in the street: *"I wouldn't feel comfortable (to use it) on the street. because it would take away from my awareness of my surrounding"*

Verification also shaped localization practices. While some participants confirmed responses by physically moving toward and touching objects, some other described a broader trial-and-error process. For example, after using her cane to locate an exit door, P8 reflected: *"It's all trial and error... you have to use your cane and problem-solving skills to figure out, hey, did the app tell me correctly, did I do something wrong?"* She described this as a collaborative effort: *"It's a tag team effort, you know, me and the app are working on this together. It's my brain and the AI."* Rather than expecting perfect accuracy, participants saw the system as a partner in iterative problem solving.

4.4.2 Description. **Trust** in description tasks was generally tied to task and condition. Participants trusted static, factual details (P6: room layout; P5: when the picture was clear). P7 found it robust across different lighting conditions. Participants were more cautious with nuanced or subjective judgments. For instance P6 stated that for measurements she *"couldn't rely on it"*. P11 trusted big-scene descriptions and explained: *"I feel like I can trust it to tell me. The big scene, like if a car in my driveway"*. P10 stated that color descriptions were not trusted at first try and sometimes requeried for consistency: *"for colors, I would probably ask it a couple of times to make sure. I keep getting the same response. especially with, blacks, grays, and navies stuff like that."*.

Verification in description tasks often involved **cross-checking with other AI tools**. For example, participants compared answers across apps when accuracy was important. One prompting approach that could support this is **ensembling**, where multiple prompts are used to solve the same problem and responses are aggregated into a final output. More broadly, participants' practices highlight opportunities for systems to provide built-in verification, either by running additional checks in the background or by surfacing alternative answers, as emerging general AI tools now do.

4.4.3 Identification. **Trust and verification** in identification tasks was closely tied to stakes. For low-stakes scenarios, confidence was high and concise or binary answers were preferred. P3 trusted it for food and clothing, P7 for classification tasks like cars, P10 for road signs or sidewalk clearance, and P5 assumed the app would only report what was on the package, not invent details. For high-stakes items, especially medication, participants deferred to sighted assistance rather than risk error (P3, P7). Confidence increased when images were clear (P5) or when repeated queries produced consistent results (P10). Some novices displayed a **priori trust** in the system, as with P8, who was using this kind of VLM based VQA system for the first time. She explained: *"I don't have a problem trusting AI. especially because of the results that I'm seeing. I think that AI just needs to remain very brutally honest, because that's what's going to make it maintain a reputable reputation."* This echoes prior reports of high self-reported trust and acceptance without verification in early interactions with multimodal LLM visual assistants, before users develop awareness of common error patterns [28, 79].

4.4.4 Reading text. How participants expressed **trust** in these scenarios was mixed: P10 trusted OCR, P11 did not. P5 trusted package text under the assumption that *"the app would not invent information."* which can tie to **priori trust** mentioned in description tasks. **Verification** in text reading tasks was less common compared to localization or description. Participants generally trusted textual outputs if they aligned with expectations. But in high-stakes cases, such as medication labels, test dates, or medical reports, they preferred sighted confirmation or cross-verifying across multiple AI tools rather than relying solely on one AI.

5 Discussion

This study provides an in-depth look at how blind users personalize and control a VLM-based conversational VQA system in both lab and real-world settings. By analyzing 418 interactions, plus participant reflections, and post-study interviews, we highlight how users actively shape system outputs, adapting prompts, verifying responses, and expressing preferences for verbosity and task specific guidance. In this section we situate our contributions within broader developments in the field.

5.1 Implications

5.1.1 Inclusive Datasets. Existing assistive VQA datasets have laid important foundations by grounding research in real images captured by blind and low-vision users. For example, VizWiz [40] provides data on short, single-turn question–answer interactions, and followup work has extended this line with visually grounding answers [27] and long-form responses that may contain information beyond the question answer such as explanations and suggestions [49]. However, these datasets are still primarily limited to one-image, one-question interactions. Recent work has begun to explore multi-turn interactions relevant to assistive contexts, such as camera guidance when information in a single image is insufficient [65], but they do so without releasing multi-turn conversations grounded in images taken by blind people in real world context.

Our dataset extends this space by providing multi-turn conversational VQA collected with 11 blind participants using *Be My AI*, spanning both controlled lab sessions and real-world use. It includes user-captured images, dialog transcripts, context-of-use metadata, and task success. This enables analysis of follow-ups, referential disambiguation, and perceptions of helpfulness in assistive scenarios. In doing so, it bridges assistive VQA datasets with the broader multi-turn visually grounded dialogue research, providing a blind-centered testbed for exploring emerging interactive behaviors.

5.1.2 Relevance to Emerging Live VQA. Our findings on image-based VQA personalization offer insights that can inform the development of the next generation of assistive video-based or live VQA systems. Recent efforts such as ChatGPT's Advanced Voice with Video for real-time assistance to blind users [24], the EgoBlind dataset of egocentric video QA [97], and prototype systems like VocalEyes [25] and AIris [20] show growing interest in real-time assistance for blind users. However, their continuous frames and faster-paced interaction, heighten, rather than reduce, the need for accurate, user-tailored, and concise responses. The customization practices we document, such as toggling verbosity under time pressure, grounding spatial descriptions in the user's perspective, offering proactive capture guidance, and supporting verification

for high-stakes tasks, can inform and support the design of more effective live QA systems.

5.1.3 Toward Teachable Generative AI. Our findings indicate a strong interest among blind participants in persistent, user-driven personalization. Many wanted the system to remember preferred detail levels (e.g., fine-grained color shades or specific cooking cues) and to consistently provide task-oriented guidance across sessions. Existing tools with memory features, such as ChatGPT and Claude, can store and recall textual preferences, but our results suggest extending these capabilities beyond text to multimodal input. Building on earlier teachable machine concepts [53, 87, 95], users could label household items once, and have those preferences persist, supporting user-driven concept learning, automatic conditioning on verbosity or ordering of information, and adaptation to new environments while retaining individual guidance. While prototypes of memory-enabled LLMs exist, no widely deployed system yet offers blind users a seamless multimodal way to *"teach"* generative AI, pointing to a direction for future assistive technology.

5.2 Limitations

Several factors related to the platform used for data collection shape the limitations of this work. All interactions were conducted using Be My AI at a fixed stage of its development. We selected this application because, at the time of the study, it was the only screen-reader–accessible, real-world VQA tool suitable for deployment with blind participants. While Be My AI was not designed for research or for persistent customization, its accessibility and widespread use made it an appropriate platform for observing authentic interaction patterns. Because the system did not retain user preferences across sessions, participants' adaptations could only influence the current dialog.

To help surface the kinds of stable preferences participants wished an ideal system would remember, we asked them to use customization techniques as if persistence were available. This approach aligns with prior work showing that effective personalization depends on a system's ability to store and reuse user-relevant information over time [77, 106]. The lack of such persistence in our deployment context likely shaped how participants experimented with customization during the diary study. For instance, P9 frequently asked about color, marbling, and marinade distribution when photographing meat—preferences she hoped a future system could remember: *"If I said the first time that I want marbling and color, then ideally it should remember that for the next time."* Such cases illustrate how participants adapted to the system's current capabilities while articulating expectations for future, more persistent forms of customization.

These platform characteristics also informed our study duration. Extending the deployment substantially longer would likely have increased fatigue, as participants would need to repeat the same preferences without any system retention. A ten-day period was deemed long enough for recurring scenarios and customization opportunities to emerge while limiting this potential burden.

6 Conclusion

In this study, we investigate how blind users exert control over AI-generated responses in conversational VQA systems through a range of prompting-based customization techniques. Through a three-phase study combining guided in-lab session and in-the-wild diary session, we find that short participant inputs elicit responses over ten times longer and can extend up to 21 turns, especially when users recapture images, clarify ambiguous outputs, or follow curiosity-driven questions. Participants actively shaped responses using both introduced and self-invented customization techniques, to manage verbosity, focus on task-relevant details, and strengthen trust. These strategies appeared across professional, familiar, and unfamiliar contexts, with slightly higher rates in familiar settings. Reflections highlight key design needs for future assistive VQA systems: on-demand verbosity control, user-centered spatial references, proactive camera guidance, and verification mechanisms for high-stakes tasks. We publicly release Ask&Prompt, our dataset capturing blind users' real-world and in-lab interactions with Be My AI, along with their reflections and detailed context of use. Together, these contributions point toward a future of assistive AI that strengthens blind users' autonomy by enabling richer customization, supporting individual goals, preferences, and everyday contexts.

Acknowledgments

The contents of this paper were developed under grant from the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR Grant No. 90REGE0024). NIDILRR is a Center within the Administration for Community Living (ACL), Department of Health and Human Services (HHS). The contents of this paper do not necessarily represent the policy of NIDILRR, ACL, HHS, and you should not assume endorsement by the Federal Government. This material is based upon work partially supported by the NSF under Grant No. 2229885 (NSF Institute for Trustworthy AI in Law and Society, TRAILS). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] 2026. National Federation of the Blind. https://nfb.org/. https://nfb.org/ Accessed: 2026-01-26.

[2] Rudaiba Adnin and Maitraye Das. 2024. "I look at it as the king of knowledge": How Blind People Use and Understand Generative AI Tools. 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 64, 14 pages. doi:10.1145/3663548.3675631

[3] Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (Athens, Greece) *(EACL '09)*. Association for Computational Linguistics, USA, 33–41.

[4] Zahra Ahmadi, Peter R. Lewis, and M. Sukhai. 2024. Reporting Risks in AI-based Assistive Technology Research: A Systematic Review. doi:10.48550/arXiv.2407.12035

[5] Aira Tech Corp. [n. d.]. Aira. https://aira.io/. Accessed: 2025-08-26.

[6] Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. 2024. MyVLM: Personalizing VLMs for User-Specific Queries. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XIII* (Milan, Italy). Springer-Verlag, Berlin, Heidelberg, 73–91. doi:10.1007/978-3-031-72624-8_5

[7] Mrim M. Alnfiai and Muhammad Ashad Kabir. 2024. Exploring the impact of word prediction assistive features on smartphone keyboards for blind users. *Heliyon* 10 (2024). https://api.semanticscholar.org/CorpusID:271909760

[8] Reem Abdulmohsen Alqahtani, Reema Saeed Alqahtani, Qamra Abdulrhman Almutairi, Reema Mohammad Aloqayli, and Haifa Hamad Alkasem. 2023. Ayn: Assistive Tool to Describe Images for Blind and Visually Impaired. In *2023 Sixth*

International Conference of Women in Data Science at Prince Sultan University (WiDS PSU). 25–30. doi:10.1109/WiDS-PSU57071.2023.00018

[9] Batyr Arystanbekov, Askat Kuzdeuov, Shakhizat Nurgaliyev, and H. A. Varol. 2023. Image Captioning for the Visually Impaired and Blind: A Recipe for Low-Resource Languages. 4 pages. doi:10.1109/EMBC40787.2023.10340575

[10] Seungho Baek, Hyerin Im, Jiseung Ryu, Juhyeong Park, and T. Lee. 2023. PromptCrafter: Crafting Text-to-Image Prompt through Mixed-Initiative Dialogue with LLM. doi:10.48550/arXiv.2307.08985

[11] Mirza Samad Ahmed Baig, Syeda Anshrah Gillani, Shahid Munir Shah, Mahmoud Aljawarneh, Abdul Akbar Khan, and Muhammad Hamzah Siddiqui. 2024. AI-based Wearable Vision Assistance System for the Visually Impaired: Integrating Real-Time Object Recognition and Contextual Understanding Using Large Vision-Language Models. arXiv:2412.20059 [cs.CV] https://arxiv.org/abs/2412.20059

[12] Advait Bal, P Anjana, K T Gayathri Devi, Vishaal S Upot, and T Anjali. 2024. Exploring the Integration of GPT-Powered Navigation and Book Reader Glasses for the Visually Impaired. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. 1–5. doi:10.1109/ICCCNT61001.2024.10725758

[13] Oliver Bendel. 2024. How Can Generative AI Enhance the Well-being of Blind? arXiv:2402.07919 [cs.HC] https://arxiv.org/abs/2402.07919

[14] 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

[15] Jeffrey P. Bigham, Chandrika Jayant, Andrew Miller, Brandyn White, and Tom Yeh. 2010. VizWiz::LocateIt - enabling blind people to locate objects in their environment. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. 65–72. doi:10.1109/CVPRW.2010.5543821

[16] Carter Blair, Kate Larson, and Edith Law. 2024. Democratizing Reward Design for Personal and Representative Value-Alignment. (2024). arXiv:2410.22203 [cs.AI] https://arxiv.org/abs/2410.22203

[17] Stephen Brade, Bryan Wang, Maurício Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-Image Generation through Interactive Prompt Exploration with Large Language Models. doi:10.1145/3586183.3606725

[18] Erin L. Brady, Yu Zhong, Meredith Ringel Morris, and Jeffrey P. Bigham. 2013. Investigating the appropriateness of social network question asking as a resource for blind users. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, Texas, USA) *(CSCW '13)*. Association for Computing Machinery, New York, NY, USA, 1225–1236. doi:10.1145/2441776.2441915

[19] Danielle Bragg, Katharina Reinecke, and Richard E. Ladner. 2021. Expanding a Large Inclusive Study of Human Listening Rates. *ACM Trans. Access. Comput.* 14, 3, Article 12 (July 2021), 26 pages. doi:10.1145/3461700

[20] Dionysia Danai Brilli, Evangelos Georgaras, Stefania Tsilivaki, Nikos Melanitis, and Konstantina Nikita. 2024. AIris: An AI-powered Wearable Assistive Device for the Visually Impaired. arXiv:2405.07606 [cs.HC] https://arxiv.org/abs/2405.07606

[21] Michele A. Burton, Erin Brady, Robin Brewer, Callie Neylan, Jeffrey P. Bigham, and Amy Hurst. 2012. Crowdsourcing subjective fashion advice using VizWiz: challenges and opportunities. 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, 135–142. doi:10.1145/2384916.2384941

[22] Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024. Personalized Steering of Large Language Models: Versatile Steering Vectors Through Bi-directional Preference Optimization. doi:10.48550/arXiv.2406.00045

[23] Ruei-Che Chang, Yuxuan Liu, and Anhong Guo. 2024. WorldScribe: Towards Context-Aware Live Visual Descriptions. doi:10.1145/3654777.3676375

[24] Ruei-Che Chang, Rosiana Natalie, Wenqian Xu, Jovan Zheng Feng Yap, and Anhong Guo. 2025. Probing the Gaps in ChatGPT Live Video Chat for Real-World Assistance for People who are Blind or Visually Impaired. arXiv:2508.03651 [cs.HC] https://arxiv.org/abs/2508.03651

[25] Kunal Chavan, Keertan Balaji, Spoorti Barigidad, and Samba Raju Chiluveru. 2024. VocalEyes: Enhancing Environmental Perception for the Visually Impaired through Vision-Language Models and Distance-Aware Object Detection. In *2024 IEEE Conference on Engineering Informatics (ICEI)*. IEEE, 1–6. doi:10.1109/icei64305.2024.10912415

[26] Khansa Chemnad and Achraf Othman. 2024. Digital accessibility in the era of artificial intelligence—Bibliometric analysis and systematic review. doi:10.3389/frai.2024.1349668

[27] Chongyan Chen, Samreen Anjum, and Danna Gurari. 2022. Grounding Answers for Visual Questions Asked by Visually Impaired People. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19076–19085. doi:10.1109/CVPR52688.2022.01851

[28] Meng Chen, Akhil Iyer, and Amy Pavel. 2025. Surfacing Variations to Calibrate Perceived Reliability of MLLM-generated Image Descriptions. doi:10.1145/3663547.3746393

[29] John Joon Young Chung and Eytan Adar. 2023. PromptPaint: Steering Text-to-Image Generation Through Paint Medium-like Interactions. doi:10.1145/3586183.3606777

[30] Be My Eyes. 2015. Be My Eyes. https://www.bemyeyes.com. Accessed: 2025-04-10.

[31] Be My Eyes. 2023. Terms of Service. https://www.bemyeyes.com/terms-conditions/

[32] Yingchaojie Feng, Xingbo Wang, Kamkwai Wong, Sijia Wang, Yuhong Lu, Minfeng Zhu, Baicheng Wang, and Wei Chen. 2023. PromptMagician: Interactive Prompt Engineering for Text-to-Image Creation. 295-305 pages. doi:10.1109/TVCG.2023.3327168

[33] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods* 5, 1 (2006), 80–92. doi:10.1177/160940690600500107 arXiv:https://doi.org/10.1177/160940690600500107

[34] Daniel Louzada Fernandes, Marcos Henrique Fonseca Ribeiro, Michel Melo Silva, and Fabio Ribeiro Cerqueira. 2024. VIIDA and InViDe: computational approaches for generating and evaluating inclusive image paragraphs for the visually impaired. *Disability and Rehabilitation: Assistive Technology* 20 (2024), 1470 – 1495. https://api.semanticscholar.org/CorpusID:274655198

[35] Iason Gabriel. 2020. Artificial Intelligence, Values and Alignment. *CoRR* abs/2001.09768 (2020). arXiv:2001.09768 https://arxiv.org/abs/2001.09768

[36] Bhanuka Gamage, Thanh-Toan Do, N. Price, A. Lowery, and Kim Marriott. 2023. What do Blind and Low-Vision People Really Want from Assistive Smart Devices? Comparison of the Literature with a Focus Study. doi:10.1145/3597638.3608955

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

[38] Meagan Griffith, Brian Wentz, and Jonathan Lazar. 2023. Quantifying the Cost of Web Accessibility Barriers for Blind Users. *Interacting with Computers* 34, 6 (2023), 137–149. doi:10.1093/iwc/iwad004

[39] Yuhan Guo, Hanning Shao, Can Liu, Kai Xu, and Xiaoru Yuan. 2024. PrompTHis: Visualizing the Process and Influence of Prompt Editing during Text-to-Image Creation. doi:10.48550/arXiv.2403.09615

[40] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People. *CoRR* abs/1802.08218 (2018). arXiv:1802.08218 http://arxiv.org/abs/1802.08218

[41] Deema Abdal Hafeth, Gokul Lal, Mohammed Al-Khafajiy, Thar Baker, and Stefanos Kollias. 2023. Cloud-IoT Application for Scene Understanding in Assisted Living: Unleashing the Potential of Image Captioning and Large Language Model (ChatGPT). In *2023 16th International Conference on Developments in eSystems Engineering (DeSE)*. 150–155. doi:10.1109/DeSE60595.2023.10468995

[42] Sherzod Hakimov and David Schlangen. 2023. Images in language space: Exploring the suitability of large language models for vision & language tasks. *arXiv preprint arXiv:2305.13782* (2023).

[43] Tracy G Harwood and Tony Garry. 2003. An overview of content analysis. *The marketing review* 3, 4 (2003), 479–498.

[44] Julia Dorothea Hayh, Siny Joseph, Susanne Dirks, and Vinod Namboodiri. 2025. Technology and trust: A scoping review on AI-enabled orientation and mobility technologies for blind and visually impaired individuals. doi:10.1177/10554181251355426

[45] Heidi Ahmed Holiel, S. Fawzi, and Walid Al-Atabany. 2024. Assisting Visually Impaired Subjects Using Large Language Models: A Comprehensive Evaluation. 561-566 pages. doi:10.1109/NILES63360.2024.10753262

[46] Jonggi Hong and Hernisa Kacorri. 2024. Understanding How Blind Users Handle Object Recognition Errors: Strategies and Challenges. 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 63, 15 pages. doi:10.1145/3663548.3675635

[47] Jonggi Hong, Christine Vaing, Hernisa Kacorri, and Leah Findlater. 2020. Reviewing Speech Input with Audio: Differences between Blind and Sighted Users. *ACM Trans. Access. Comput.* 13, 1, Article 2 (April 2020), 28 pages. doi:10.1145/3382039

[48] A Huberman et al. 2014. *Qualitative data analysis a methods sourcebook*. Thousand Oaks, Californinia SAGE Publications, Inc.

[49] Mina Huh, Fangyuan Xu, Yi-Hao Peng, Chongyan Chen, Hansika Murugu, Danna Gurari, Eunsol Choi, and Amy Pavel. 2024. Long-Form Answers to Visual Questions from Blind and Low Vision People. arXiv:2408.06303 [cs.CL] https://arxiv.org/abs/2408.06303

[50] EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning Language Models to User Opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 5906–5919. doi:10.18653/v1/2023.findings-emnlp.393

[51] Mohamed Fazli Imam, Chenyang Lyu, and Alham Fikri Aji. 2025. Can Multimodal LLMs do Visual Temporal Understanding and Reasoning? The answer is No! *arXiv preprint arXiv:2501.10674* (2025).

[52] Md Touhidul Islam, Imran Kabir, Elena Ariel Pearce, Md Alimoor Reza, and Syed Masum Billah. 2024. Identifying Crucial Objects in Blind and Low-Vision Individuals' Navigation. In *The 26th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '24)*. ACM, 1–8. doi:10.1145/3663548.3688538

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

[54] Antonia Karamolegkou, Malvina Nikandrou, Georgios Pantazopoulos, Danae Sanchez Villegas, Phillip Rust, Ruchira Dhar, Daniel Hershcovich, and Anders Søgaard. 2025. Evaluating Multimodal Language Models as Visual Assistants for Visually Impaired Users. arXiv:2503.22610 [cs.HC] https://arxiv.org/abs/2503.22610

[55] Jee-Eun Kim, Gurung Sahas, and Masahiro Bessho. 2025. Toward Assisting Blind Individuals in Exploring Unfamiliar Indoor Environments Using Multimodal LLM and Smartphone LiDAR. In *2025 IEEE International Conference on Consumer Electronics (ICCE)*. 1–6. doi:10.1109/ICCE63647.2025.10929822

[56] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence* 6, 4 (April 2024), 383–392. doi:10.1038/s42256-024-00820-y

[57] Dimitri Lacroix, Ricarda Wullenkord, and Friederike Eyssel. 2023. Who's in Charge?: Using Personalization vs. Customization Distinction to Inform HRI Research on Adaptation to Users. *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (2023). https://api.semanticscholar.org/CorpusID:257406282

[58] Walter S. Lasecki, Phyo Thiha, Yu Zhong, Erin Brady, and Jeffrey P. Bigham. 2013. Answering visual questions with conversational crowd assistants. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility* (Bellevue, Washington) *(ASSETS '13)*. Association for Computing Machinery, New York, NY, USA, Article 18, 8 pages. doi:10.1145/2513383.2517033

[59] Tung Le, Huy Tien Nguyen, and Minh Le Nguyen. 2021. Multi visual and textual embedding on visual question answering for blind people. *Neurocomputing* 465 (2021), 451–464. doi:10.1016/j.neucom.2021.08.117

[60] 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

[61] Young-Jun Lee, Chae-Gyun Lim, Yunsu Choi, Ji-Hui Lm, and Ho-Jin Choi. 2022. PERSONACHATGEN: Generating Personalized Dialogues using GPT-3. In *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge*, Heuiseok Lim, Seungryong Kim, Yeonsoo Lee, Steve Lin, Paul Hongsuck Seo, Yumin Suh, Yoonna Jang, Jungwoo Lim, Yuna Hur, and Suhyune Son (Eds.). Association for Computational Linguistics, Gyeongju, Republic of Korea, 29–48. https://aclanthology.org/2022.ccgpk-1.4/

[62] Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large Language Models Understand and Can be Enhanced by Emotional Stimuli. (2023). arXiv:2307.11760 [cs.CL] https://arxiv.org/abs/2307.11760

[63] Junyi Li, Charith Peris, Ninareh Mehrabi, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2024. The steerability of large language models toward data-driven personas. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 7290–7305. doi:10.18653/v1/2024.naacl-long.405

[64] Jiongnan Liu, Zhicheng Dou, Guoyu Tang, and Sulong Xu. 2023. JDsearch: A Personalized Product Search Dataset with Real Queries and Full Interactions. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2945–2952. doi:10.1145/3539618.3591900

[65] Li Liu, Diji Yang, Sijia Zhong, Kalyana Suma Sree Tholeti, Lei Ding, Yi Zhang, and Leilani H. Gilpin. 2024. Right this way: Can VLMs Guide Us to See More to Answer Questions? arXiv:2411.00394 [cs.CV] https://arxiv.org/abs/2411.00394

[66] Albert Lu, Hongxin Zhang, Yanzhe Zhang, Xuezhi Wang, and Diyi Yang. 2023. Bounding the Capabilities of Large Language Models in Open Text Generation with Prompt Constraints. arXiv:2302.09185 [cs.CL] https://arxiv.org/abs/2302.09185

[67] Kaixiang Mo, Yu Zhang, Shuangyin Li, Jiajun Li, and Qiang Yang. 2018. Personalizing a dialogue system with transfer reinforcement learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence* (New Orleans, Louisiana, USA) *(AAAI'18/IAAI'18/EAAI'18)*. AAAI Press, Article 652, 8 pages.

[68] Ananya Gubbi Mohanbabu and Amy Pavel. 2024. Context-Aware Image Descriptions for Web Accessibility. doi:10.1145/3663548.3675658

[69] S Shiv Mohith, S Vijay, Sanjana V, and Niranjana Krupa. 2020. Visual World to an Audible Experience: Visual Assistance for the Blind And Visually Impaired. *2020 IEEE 17th India Council International Conference (INDICON)* (2020), 1–6. https://api.semanticscholar.org/CorpusID:231852686

[70] Francisco Willem R. Moreira, Guilherme Hermes, and Jean Mário M. de Lima. 2024. Development of a Cross Platform Mobile Application Using Gemini to Assist Visually Impaired Individuals. 559-566 pages. doi:10.1109/ICIIBMS62405.2024.10792854

[71] Meredith Ringel Morris, Jazette Johnson, Cynthia L. Bennett, and Edward Cutrell. 2018. Rich Representations of Visual Content for Screen Reader Users. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3173574.3173633

[72] Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. 2024. Yo'LLaVA: Your Personalized Language and Vision Assistant. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=mjGy8g3pgi

[73] Minhyeon Oh, Seungjoon Lee, and Jungseul Ok. 2024. Active Preference-based Learning for Multi-dimensional Personalization. doi:10.48550/arXiv.2411.00524

[74] OpenAI. 2024. Memory and New Controls for ChatGPT. https://openai.com/index/memory-and-new-controls-for-chatgpt/. Accessed: 2025-04-10.

[75] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 2011, 15 pages.

[76] Hyerim Park, Malin Eiband, André Luckow, and Michael Sedlmair. 2025. Exploring Visual Prompts: Refining Images with Scribbles and Annotations in Generative AI Image Tools.

[77] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442 [cs.HC] https://arxiv.org/abs/2304.03442

[78] Pruthvi Patel, Swaroop Mishra, Mihir Parmar, and Chitta Baral. 2022. Is a Question Decomposition Unit All We Need?. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 4553–4569. doi:10.18653/v1/2022.emnlp-main.302

[79] Ricardo E. Gonzalez Penuela, Ruiying Hu, Sharon Y. Lin, Tanisha Shende, and Shiri Azenkot. 2025. Towards Understanding the Use of MLLM-Enabled Applications for Visual Interpretation by Blind and Low Vision People. doi:10.1145/3706599.3719714

[80] PiccyBot. 2024. PiccyBot. https://www.piccybot.com/. Accessed: 2025-04-10.

[81] Shrinivas J. Pundlik, Prerana Shivshanker, and Gang Luo. 2023. Impact of Apps as Assistive Devices for Visually Impaired Persons. doi:10.1146/annurev-vision-111022-123837

[82] Anjana S, Arya J, Sandra Jayaraj, Sreelakshmi Raj, and Sruthy Manmadhan. 2024. VisQAid: Visual Question Answering for Visually Impaired. 1–6. doi:10.1109/RAICS61201.2024.10690148

[83] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When Large Language Models Meet Personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7370–7392. doi:10.18653/v1/2024.acl-long.399

[84] Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Megan L. Rogers, Inna Goncearenco, Giuseppe Sarli, Igor Galynker, Denis Peskoff, Marine Carpuat, Jules White, Shyamal Anadkat, Alexander Hoyle, and Philip Resnik. 2025. The Prompt Report: A Systematic Survey of Prompt Engineering Techniques. arXiv:2406.06608 [cs.CL] https://arxiv.org/abs/2406.06608

[85] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA Models That Can Read. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*

(CVPR) (2019), 8309–8318. https://api.semanticscholar.org/CorpusID:85553602
[86] Sangmim Song, S. Kodagoda, A. Gunatilake, Marc. G Carmichael, K. Thiyagarajan, and Jodi Martin. 2024. Guide-LLM: An Embodied LLM Agent and Text-Based Topological Map for Robotic Guidance of People with Visual Impairments. doi:10.48550/arXiv.2410.20666
[87] Joan Sosa-García and Francesca Odone. 2017. "Hands On" Visual Recognition for Visually Impaired Users. *ACM Trans. Access. Comput.* 10, 3, Article 8 (Aug. 2017), 30 pages. doi:10.1145/3060056
[88] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376404
[89] Jian Tang, Yi Zhu, Gai Jiang, Lin Xiao, Wei Ren, Yu Zhou, Qinying Gu, Biao Yan, Jiayi Zhang, Hengchang Bi, Xing Wu, Zhiyong Fan, and Leilei Gu. 2025. Human-centred design and fabrication of a wearable multimodal visual assistance system. *Nature Machine Intelligence* (April 2025). doi:10.1038/s42256-025-01018-6
[90] Mohammad Moeen Valipoor and A. de Antonio. 2022. Recent trends in computer vision-driven scene understanding for VI/blind users: a systematic mapping. 983-1005 pages. doi:10.1007/s10209-022-00868-w
[91] Ehsan Variani, Khe Chai Sim, Kilol Gupta, Lara McConnaughey, Mingqing Chen, Rajiv Mathews, Shefali Garg, Swaroop Ramaswamy, and Theresa Breiner (Eds.). 2022. *UserLibri: A Dataset for ASR Personalization with Only Text.* Accepted to Interspeech 2022.
[92] Zhijie Wang, Yuheng Huang, Da Song, Lei Ma, and Tianyi Zhang. 2024. PromptCharm: Text-to-Image Generation through Multi-modal Prompting and Refinement. doi:10.1145/3613904.3642803
[93] Eleanor Watson, Minh Nguyen, Sarah Pan, and Shujun Zhang. 2025. Choice Vectors: Streamlining Personal AI Alignment Through Binary Selection. doi:10.3390/mti9030022
[94] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1800, 14 pages.
[95] Linda Yilin Wen, Cecily Morrison, Martin Grayson, Rita Faia Marques, Daniela Massiceti, Camilla Longden, and Edward Cutrell. 2024. Find My Things: Personalized Accessibility through Teachable AI for People who are Blind or Low Vision. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 403, 6 pages. doi:10.1145/3613905.3648641
[96] Barbara H. Wixom and Peter A. Todd. 2005. A Theoretical Integration of User Satisfaction and Technology Acceptance. *Information Systems Research* 16, 1 (2005), 85–102. doi:10.1287/isre.1050.0042
[97] Junbin Xiao, Nanxin Huang, Hao Qiu, Zhulin Tao, Xun Yang, Richang Hong, Meng Wang, and Angela Yao. 2025. EgoBlind: Towards Egocentric Visual Assistance for the Blind. arXiv:2503.08221 [cs.CV] https://arxiv.org/abs/2503.08221
[98] Jingyi Xie, Rui Yu, He Zhang, Sooyeon Lee, Syed Masum Billah, and John M. Carroll. 2024. Emerging Practices for Large Multimodal Model (LMM) Assistance for People with Visual Impairments: Implications for Design. arXiv:2407.08882 [cs.HC] https://arxiv.org/abs/2407.08882
[99] Rui Yan, Jian-Yun Nie, and Xiaoming Li. 2011. Summarize What You Are Interested In: An Optimization Framework for Interactive Personalized Summarization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Regina Barzilay and Mark Johnson (Eds.). Association for Computational Linguistics, Edinburgh, Scotland, UK., 1342–1351. https://aclanthology.org/D11-1124/
[100] Joy Yun, Yann Hicke, Mariah Olson, and Dorottya Demszky. 2024. Enhancing Tutoring Effectiveness Through Automated Feedback: Preliminary Findings from a Pilot Randomized Controlled Trial on SAT Tutoring. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (Atlanta, GA, USA) *(L@S '24)*. Association for Computing Machinery, New York, NY, USA, 422–426. doi:10.1145/3657604.3664681
[101] He Zhang, Nicholas J. Falletta, Jingyi Xie, Rui Yu, Sooyeon Lee, Syed Masum Billah, and John M. Carroll. 2024. Enhancing the Travel Experience for People with Visual Impairments through Multimodal Interaction: NaviGPT, A Real-Time AI-Driven Mobile Navigation System. arXiv:2410.04005 [cs.HC] https://arxiv.org/abs/2410.04005
[102] Jianfei Zhang, Jun Bai, Bei Li, Yanmeng Wang, Rumei Li, Chenghua Lin, and Wenge Rong. 2024. Disentangling Preference Representation and Text Generation for Efficient Individual Preference Alignment. doi:10.48550/arXiv.2412.20834
[103] Sen Zhang, Jian Zeng, Xiaoyang Liu, and Siyuan Ding. 2019. Effect of obstacle density on the travel time of the visually impaired people. *Fire and Materials* 43, 2 (2019), 162–168. doi:10.1002/fam.2681 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/fam.2681
[104] Siyan Zhao, John Dang, and Aditya Grover. 2024. Group Preference Optimization: Few-Shot Alignment of Large Language Models. (2024). arXiv:2310.11523 [cs.LG] https://arxiv.org/abs/2310.11523
[105] Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized Dialogue Generation with Diversified Traits. *ArXiv* abs/1901.09672 (2019). https://api.semanticscholar.org/CorpusID:59316441
[106] Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. MemoryBank: enhancing large language models with long-term memory. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'24/IAAI'24/EAAI'24)*. AAAI Press, Article 2198, 8 pages. doi:10.1609/aaai.v38i17.29946
[107] Dong Zhou, Séamus Lawless, Xuan Wu, Wenyu Zhao, and Jianxun Liu. 2016. Enhanced Personalized Search using Social Data. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Jian Su, Kevin Duh, and Xavier Carreras (Eds.). Association for Computational Linguistics, Austin, Texas, 700–710. doi:10.18653/v1/D16-1067
[108] Jin Peng Zhou, Katie Z Luo, Jingwen Gu, Jason Yuan, Kilian Q. Weinberger, and Wen Sun. 2024. Orchestrating LLMs with Different Personalizations. doi:10.48550/arXiv.2407.04181

A Introducing Customization Techniques in the Lab

Scenario 1: Computer Screen Error Message - No Technique. The first scenario task is using the system to interpret an error message on a laptop screen. This read-text task appears frequently in the VizWiz dataset and was validated by the pilot researcher as a task he would personally use *Be My AI* for. This initial task serves to familiarize participants with the system's core functionality while giving them a chance to practice sharing their interaction session with the research team. To avoid cognitive overload, we do not introduce any customization techniques during this task.

Scenario 2: Exit Door Navigation - Binary Feedback and Style Prompting. In the second scenario, participants are asked to navigate to an exit door in an unfamiliar hallway using the system. Many studies have investigated approaches to assist blind individuals with navigation tasks [e.g. 52, 89]. After completing the interaction, participants are introduced to the binary feedback technique, which they can use to provide simple responses–such as *"helpful"* or *"not helpful"*–to indicate their preferences. This technique is framed as a way to signal to the system what kinds of responses they found useful, imagining that future responses might be shaped accordingly. We also introduce participants to style prompting, which enables them to explicitly guide the format or tone of the system's responses. For example, they might instruct the system to *"be less verbose"* or *"give more details when giving directions."*

Scenario 3: Nutrition Facts - Intention Prompting. In this task, participants are asked to find the nutrition information on a Lipton soup mix box. Before beginning, they are encouraged to set a specific goal for the task, such as monitoring sugar intake and determining whether the product is suitable for their diet. Tasks involving nutritional content also frequently appear in the VizWiz dataset. The technique introduced in this scenario is intention prompting, in which participants include their goal as part of their query such as *"What is the nutrition of this pacakge? I'm trying to control sugar intake."* An ideal system would use this context to tailor its response—omitting irrelevant details (e.g., calcium content) and focusing on information aligned with the user's stated intention. This approach also reduces the need for users to ask separate questions, as the system could provide all relevant details based on the intention.

Scenario 4: Item Localization - Chain of Thought and Decomposition Prompting. In the fourth scenario, participants are asked to find the same box of soup used in the previous task, now placed somewhere within the lab. This task focuses on localization and navigation. Rather than using the typical Chain-of-Thought (CoT) prompting phrase: *"Let's think step by step,"* we introduce a more natural way for participants to prompt with structured reasoning. For example, a participant can say, *"I only care about locating the soup, not the other items in the scene like the painting on the wall."* With explicit reasoning and instruction, an ideal system would be better in generating responses that align with users' informational needs. In diaries, we also observe participants naturally using a series of decomposed questions to work toward a solution (see subsection 4.2). This pattern resembles both CoT prompting and decomposition prompting. Given the overlap in purpose and function [84], we group these prompting techniques together in our analysis.

Scenario 5: Personal Item Localization - SimToM Prompting. The final scenario involves locating a personal item of the participant's choice (e.g., keys) that has been deliberately misplaced in the lab room. Locating personal items is a common real-world use case for blind users of visual assistance systems [95]. In this task, participants are introduced to SimToM (Simulate the Thought of a Mind) prompting, a technique that encourages them to guide the system by describing facts they would use to approach the task themselves. Here, participants are instructed to describe the item's features (e.g., shape, color, or attached objects) to help the system better understand what to look for. By incorporating this kind of explicit reasoning and instruction, the system is expected to effective and accurate responses.

B Diary Study Practice Days Protocol

On Day 1, participants are asked to try using the system to identify everyday objects within their home, such as food or clothing. On Day 2, they are encouraged to explore how lighting conditions may affect the system's performance by taking images in natural daylight, with poor lighting conditions, and with artificial lighting. They are also encouraged to try the system with accessibility-related objects (e.g., canes). On Day 3, participants are asked to step outside of familiar spaces and try the system in new or social settings, such as outdoors or around unfamiliar people. Participants are instructed to complete at least three interactions per day during the three-day practice session.

C End-of-Day Diary Study Reflection Questions

During the diary study, participants received a set of reflection questions at the end of each day, including: (1) Was the system helpful today? (2) Did you try to give it any feedback or make it work better for you? (3) How could you tell if the system was right or wrong? and (4) Is there anything else you would like to add? Responses to these questions, alongside the interaction logs, serve as key data sources in our analysis of real-world usage and customization techniques.

D Our categorization on the types of questions asked by participants

See table 3

Table 3: Types of questions asked by participants, with definitions and examples.

Question Types	Definition	Example Questions
Description Questions (40.38%*)	Questions asked for the description of a picture or the visual characteristics of an object or a scene.	"Is anyone sitting on the bench?" "What color are the pillows on the couch?" "Please just describe in the greatest possible detail the picture [on the screenshot]."
Reading Questions (20.33%)	Questions asked for reading text or looking for information based on text.	"I need to know who the sender is." "What is written on the door?" "May you read me the setup instructions?"
Localization Questions (9.62%)	Questions asked for the location or spatial relation of objects.	"Where is the ID card?" "Help me find the remote." "From where I am, how far is the dumpster and is it to my left or right?"
Identification Questions (8.79%)	Questions asked for identifying an object by name or type[18].	"What is it?" "What is in the boxes?" "What's the Stonyfield's product? Is it yogurt"
Unanswerable Questions (5.49% unanswerable till the end)	Questions that cannot be answered based on the information included in previous images in the interaction.	–

*Percentage is calculated from all diary study interactions that include questions of this type.