A Web-Based Guided Selfie Assistant for People with Visual Impairments

Yihang Wang
ywang920@jh.edu
Johns Hopkins University
Baltimore, Maryland, USA

Anubhav De
ade11@jh.edu
Johns Hopkins University
Baltimore, Maryland, USA

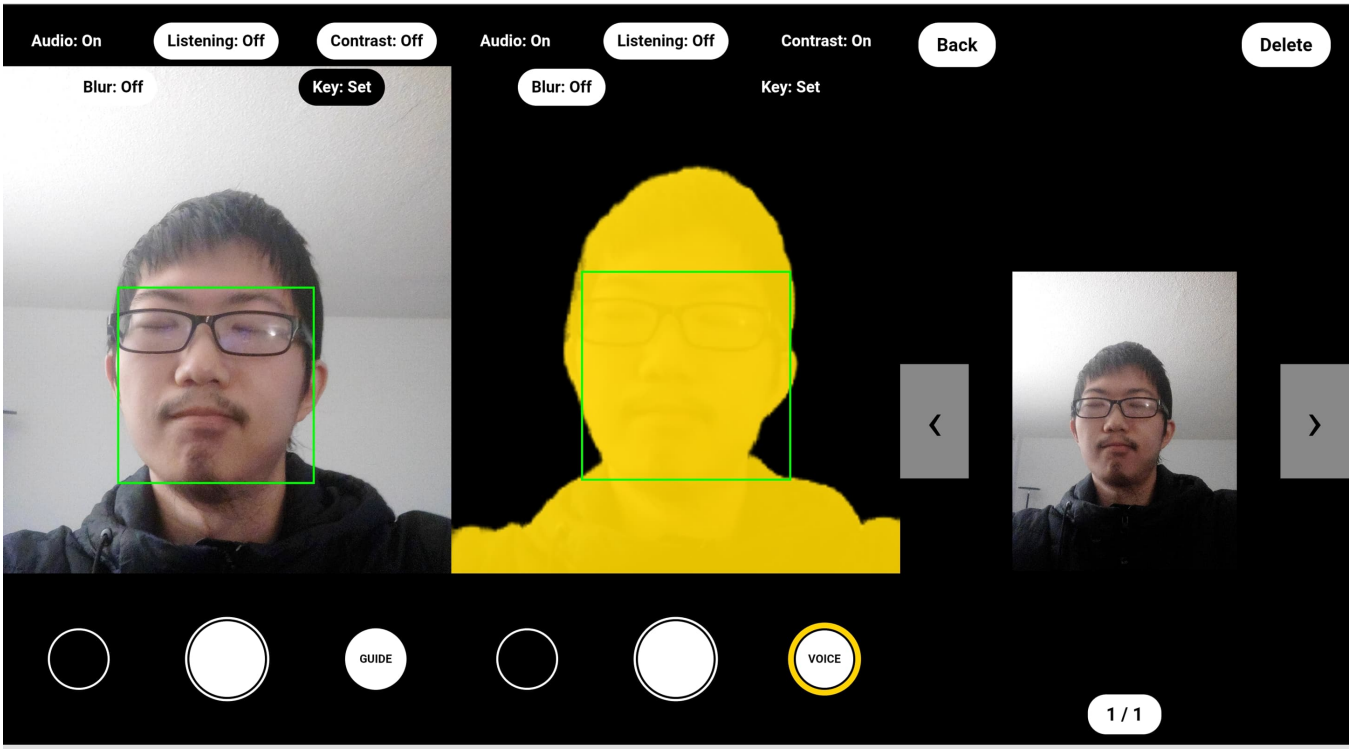

Figure 1: Overview of the Web-Based Selfie System Interface

Abstract

People with visual impairments commonly have difficulties in taking selfies because it is hard to properly frame themselves within the camera display. Whereas existing tools, such as VoiceOver and Guided Frame, provide basic directional cues, most often they rely on non-contextual and non-conversational feedback, depend on vibration cues that are undesirable for many PVI, and do not allow the human-like guiding that users prefer. Motivated by prior research showing that PVI consider selfies an important mode for self-expression, we designed and implemented a web-based, conversational guided selfie assistant that aims at supporting independent and expressive selfie-taking.

Our system combines real-time camera analysis with a flexible natural-language interface, allowing users to ask open-ended questions, e.g., "How does my framing look?" or specify personalized success criteria for their photos. An assistant provides two forms of guidance-concise directive feedback and richer conversational descriptions-addressing core design criteria identified in recent literature. Through iterative prototyping, usability inspection, and simulated user testing, we developed the system and implemented a complete, functional end-to-end web application featuring voice interaction, multimodal guidance, a virtual album, and screen-reader-friendly components. Limitations in participant recruitment prevented formal user evaluation with blind and low-vision participants, but the resulting system demonstrates the feasibility of delivering human-like goal-oriented selfie support entirely within a browser environment.

This project demonstrates both the opportunity and challenge in developing accessible AI-driven PVI tools, and points towards future work including native mobile implementations, improved multimodal non-verbal cues, and comprehensive evaluation with real users.

,

Keywords

Accessible photography, Assistive computer vision, Blind and low-vision users, Human–computer interaction, Web-based guidance systems

1 Introduction

The selfie functionality in the camera has become a common element in the everyday digital communication through which people can represent themselves, share experiences, and participate in social spaces. Research has determined that people with visual impairments value selfies as much as the sighted people do, using them as an important medium for self-expression, identity, and autonomy. However, taking a selfie independently is still not a trivial task for many PVI because it is difficult to frame one's face or body inside the camera view. This often forces users either to depend on others or take multiple unsuccessful photos, undermining independence and reducing engagement in digital platforms.

Current accessibility tools provide only partial support for taking selfies.Systems, like camera feedback in VoiceOver and Guided Frame in Google Pixel, offer directional cues such as "move right" or "raise your phone," but their feedback is usually short, inflexible, and non-conversational. These tools seldom take into account the intent of the user, aesthetic preference, or context.They also rely heavily on vibration-based cues that some PVI find confusing, uncomfortable, or too non-descriptive. Past work-including Gonzalez Penuela et al. (2022) emphasizes the need for the selfie guidance systems that can adapt to the user-defined goals, and give meaningful descriptions, and also interact naturally and human-like. Such systems should support open-ended communication and let users themselves define what constitutes a "good" photo.

Recent developments in web technologies, multimodal models, and speech interfaces finally allow the exploration of more flexible, conversational, and context-aware approaches for selfie assistance. However, there is still a gap in accessible, lightweight systems that are easy to deploy, do not require native mobile installation, and would enable PVI to obtain rich real-time guidance directly in the browser.

To this end, we introduce a web-based guided selfie assistant that enables PVIs to take well-framed selfies on their own through natural language interaction. Our system provides real-time framing feedback, human-like descriptive guidance, and voice-based conversational controls. Users can explicitly request specific types of feedback, such as "Tell me if my face is centered"; they can ask openended questions, such as "How do I look right now?"; or they can specify personalized success criteria for the photo. Our system is designed to work completely in-browser, offering universally accessible interaction modes, ARIA-labeled components, and a simplified guidance mode modeled after existing tools for users who prefer concise instructions. We present in this report the motivation for the system, our design process informed by literature on PVI needs, the architecture of our web-based implementation, and results from our prototyping and simulated evaluation efforts. While efforts to recruit blind and low-vision participants for formal usability testing were subject to timing and logistical constraints, this end-to-end implementation illustrates the feasibility and potential of conversational AI-driven selfie guidance for PVI. We conclude with reflections on system limitations, technical challenges, and future work opportunities including native mobile deployment and expanded multimodal feedback.

2 Related Works

Photography and, more so, selfie taking have become a very significant channel for self-expression, identity construction, and social participation for the Blind or low vision people. Previous qualitative research proves that selfies are valued by BLV users as much as by sighted users by relying on them to express personality, document life events, and participate in social networks. At the same time, however, they constantly face challenges related to framing, centering, and understanding visual context [1]. Conventional smartphone camera interfaces provide very little support to address these challenges, and this gap between what BLV users want and existing systems offer drives frustration, reduced independence, and reliance on sighted assistance. The prior work of Gonzalez Penuela et al. systematically documented these challenges and emphasized how BLV users prefer contextual, richly descriptive feedback and not short and directive cues. They stressed that such assistance should be flexible enough to be able to afford aesthetic goals that vary, and ideally it should support conversation-like interaction as one may receive from a human helper.

Mobile platforms have made initial steps toward accessible selfie-taking, most notably with Google Pixel's "Guided Frame," which uses computer vision to detect faces and provides audio prompts such as "move your phone slightly right and up" along with haptic cues to aid framing [2]. While highly appreciated for the hands-free framing and automated capture it enables, the design philosophy behind Guided Frame rests on short, instructive cues rather than interactive dialog. Its interaction pattern, based on vibration, which BLV users find uncomfortable or too abstract in detail, leaves many of the subtleties they reportedly want-descriptive context, expression feedback, clarifications, or personalized framing goals-out of its reach. Guided Frame is only available on specific devices, and the strict structure of prompts it uses limits flexibility across diverse user preferences or use contexts.

Beyond purely commercial features, researchers have developed other multimodal approaches to improving photo capture accessibility. For example, Lim et al.'s TouchPhoto system introduced audio, visual, and tactile modalities to support BLV users in camera alignment and photo comprehension, showing how multimodality can mitigate the limitations of audio alone as a feedback modality [3]. Similarly, Yasmin presented research on "haptic selfies," an investigation into tactile affordances for face and appearance representation; such non-verbal cues about expression and composition complement those that one might achieve through hearing alone [4]. These contributions illustrate the benefit of diversification in employed sensory modalities to enhance overall accessibility; however, many of these prototypes require specialized hardware or native apps, which limits broader deployment. Their design implications-especially in terms of richer, more complex visual content representations-remain highly germane when the focus shifts to audio interfaces alone.

At the same time, large datasets and models developed for the blind photography domain emphasize the difficulty of interpreting

the images that are taken by BLV users. A dataset called VizWiz by Gurari et al. consists of real images taken by photographers who are blind and paired audio questions. The authors demonstrated that even state-of-the-art VQA vision-and-language models have great difficulty with common photography characteristics when performed by a BLV user such as poorly lit, extremely blurred, or subjects partially out of the frame 5. This paper is one example of a larger design lesson: systems which have only analyzed a photo after capture usually fail because the image is already compromised. Instead, BLV users need real-time feedback during framing, and that feedback needs to be so flexible that it can address such questions as "Is my face centered?" or "How close am I to the top of the frame?" before the image is taken. The limitations discovered in VizWiz motivate the development of systems that combine continuous vision feedback and conversational question support. Photography research has also considered post-capture workflows, including how BLV users edit, select, and share images. Bennett et al. discovered that when assessing photos, visually impaired teenagers frequently rely on contextual cues, memory, or trial-and-error, and that accessible feedback is necessary for independent decision-making to take place [6]. Their results strengthen the argument that capture support must go beyond framing an image to convey whether the resulting image meets the user's goals. These more general insights will inform the development of systems that can provide flexible, expressive descriptions instead of strictly binary "good/bad" feedback. Taken together, this prior work highlights several gaps that motivated our system. Commercial systems provide concise directive cues, but fail to engage users in rich, goal-driven conversation. Prototypes offer rich multimodal feedback, but remain hard to deploy due to hardware requirements. Vision-language datasets suggest that post-capture interpretation cannot suffice for BLV photography since many images captured without real-time guidance feature unfixable issues. Across these lines of work, there is a clear need for an accessible, platform-agnostic selfie assistant that incorporates continuous framing feedback with natural, human-like conversation. Our web-based guided selfie system is designed to fill this gap by leveraging modern browser capabilities and multimodal AI models to provide descriptive guidance, open-ended questions, and user-defined success criteria while remaining lightweight enough to run on any modern device without installation.

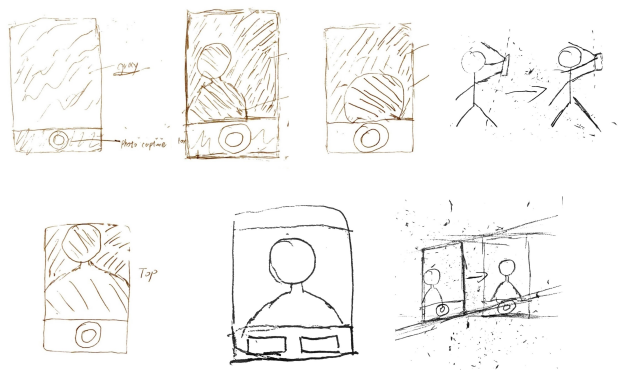

Figure 2: Low-fidelity Prototype

3 Methods

3.1 Prototype and Implementation

3.1.1 Low-fidelity Prototype. We initially designed the low-fidelity prototype based on user needs learned from previous papers on this topic, as shown in Figure 2. We were directly inspired by previous research by Gonzalez Penuela et al., which identifies several requirements for a good selfie system[1]. These include allowing users to (1) evaluate their selfie based on different criteria, (2) talk with the system like a human, (3) receive non-verbal audio feedback, and (4) modify how much information is given to them.

We chose to support all of the requirements except non-verbal audio feedback. Due to the difficulty of recruiting participants for contextual inquiries, we were unable to determine what kind of non-verbal audio feedback users would prefer in different scenarios.

We designed two different interaction flows, which corresponds to two modes that give different levels of information to the user to support (4). The first mode is Voice Control mode, where users can ask questions about the current frame/photo, or command the system to do something (like guiding them to take a photo, navigating the interface, etc.). To support (2), both of these two operations are designed to support natural language input, instead of certain phrases. By allowing users to ask anything about a selfie using natural language, we also aim to support (1), since users may want to evaluate their selfies based on criteria we can't predict. The second mode is Simple Guidance Mode, where users are continuously prompted by the system about their position and distance in the frame. For example, if the user's face is too left in the frame, the system will ask the user to turn their phone to the right. This mode is similar to what Apple's VoiceOver or Google's Guided Frame does, where verbal feedback is conveyed in one-way and a passive way. We designed this mode to simplify the process that one user needs to go through to take a simple centered selfie.

We decided to design the interface to mimic the common camera interfaces on smartphones to reduce learning costs. There is a camera view and a virtual photo album view. Users use the camera view to take selfies or ask questions about the current frame, and use the album view to review the photos previously taken.

When designing the low-fidelity prototype, we used a smartphone's selfie camera to simulate the system to validate if any interactions at least make sense to us, since the prototype was not a working system yet. It was in this process we decided to not add non-verbal audio feedback, since we could not finalize a satisfying design decision. Another thing we noticed during this was that in most cases users should adjust their position by rotating the phone instead of moving it horizontally.

3.1.2 Web Interface Overview. To enable quick development, we chose to build the system based on a web tech stack. Also, due to time constraints, we directly built the web interface from the low-fidelity prototype without first creating a high-fidelity one. The interface is shown in Figure 1. As mentioned in the prototyping section, the interface was designed to be similar to the common camera app on a smartphone, including a camera view and an album view. More details were added when building this from the prototype. The whole interface is mainly black and white to create

a high contrast. The camera view is separated into three parts. They were the top bar, the camera preview area and the bottom bar.

The top bar is where users could toggle different features. To be more specific, there are buttons that users could click to mute/unmute the audio, enable/disable the microphone, turn on/off the high-contrast and background blur effects, and set the user key. The high-contrast and background blur effects both aim to highlight the user as the main object in the frame and reduce the background's presence for low-vision users. The former displays the user in the frame as a yellow silhouette against a black background and the latter blurs the background. The captured photo only renders the blur effect. Audio and microphone buttons do what their name suggests to control the audio output and input. The key button is used to set the user key, which will be discussed later.

The bottom bar contains three main buttons in this interface. The first one is the album button. One could click it to navigate to the album view. The second one is the capture button. One could click it to take a photo instantly. The third one is a "mode switch" button. One could click it to switch between Voice Control mode and Simple Guidance mode. They are all designed to be in a circular shape to resemble the common design.

The album photo view includes several parts. The navigation bar allows the user to go back to the camera view and delete the current photo. The photo review area contains the photo that is displayed currently, including two buttons that enables the user to see the last or next photo. There is also a bottom element showing the index of the current photo.

3.1.3 Main Interaction Flows. As mentioned before, there are two different modes that users can select. One is simple Guidance mode, another one is Voice Control Mode. When in simple guidance mode, we firstly get face bounding box information from face detection service module, like face position or size in the frame. Then we normalize it and send it into an evaluation function that basically has some thresholds set to evaluate and give back evaluation results, like "Left" or "Top". It will then be converted into human understandable language using predefined matching rules. For example, when the user's face is too left, we tell them to turn the phone right. And this whole process continues to happen in a loop to guide the user to move their face to a centered and normal distance.

However, more complex user needs remain unaddressed. How can users judge their selfie using standards beyond securing a centered face? How can users express their needs just like talking with a real human? To address this, we introduce the Voice Control Mode. When in this mode, user' microphone is continuously monitored and we used VAD to detect if a real human speech happens. When the user speaks, the speech-to-text module transcribes the audio into text, which is then sent to the backend, along with a list of available tools and their descriptions. The backend then returns a response containing either text or a tool invocation, and the frontend converts the text to audio or executes the specified tool accordingly.

3.1.4 System Implementation. The whole system is composed of a web frontend and a serverless backend. The user can use a smartphone to use this app in the browser without installation. This

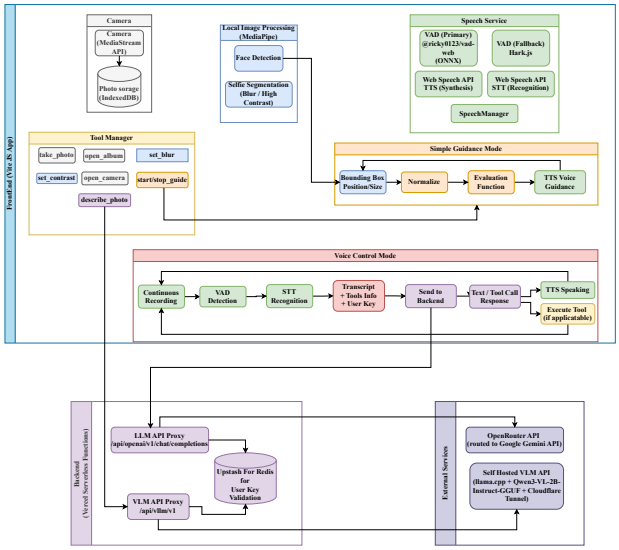

Figure 3: Software Architecture of the Web-Based Selfie System

architecture choice, as mentioned in the last section, is for quick iteration and prototyping purposes. The overview of this architecture is shown in Figure 3.

Previous work has shown that people with visual impairments highly value privacy when using camera-based assistive tools[**?**]. For this reason, we chose to make most computation tasks happen at the frontend side. We use IndexedDB to store any photos locally in the browser using IndexedDB. For simple image processing tasks, like face detection or segmentation, we also make them happen at the client side using MediaPipe. For any text to speech or speech to text tasks, we use open-source in-browser VAD(Voice activity detection) implementations (vad-web for main use and Hark.JS as fallback) plus Web Speech API (although the latter one's locality depends on user's browser implementation). Those computation tasks, combined with other simple interface interactions, are wrapped into different "tools" that can be triggered separately. How these tools are used during user interactions will be covered later in this section.

It goes without saying that the basic web accessibility matters here for the frontend part. To ensure this, we used sematic elements instead of normal ones for the web page. For example, we used <button> for any clickable elements instead of . We also added extra ARIA labels to elements to make sure that they could be determined by screen readers or other assisted tools. For instance, we added "Camera Controls" to the bottom bar to indicate its purpose. For camera preview and photo review area, we not only added ARIA labels to indicate what they are, but also dynamically update face detection results inside the labels. For a photo, it could be read out as "Camera preview, 1 face detected, face centered."

The backend is the place we moved those that are not feasible to run at the client side, like the LLM and VLM(Vision language model) API proxy. Before we finalize this backend architecture, we

Table 1: Participant Profile (Demographic)

Item	Details
ID	P1
Vision status	Low vision
Gender	Male
Smart phone use	Yes (Google Pixel)
Screen reader Use	Yes (TalkBack)
Language of communication	English

also tried other solutions like WASM(Web Assembly) or WebGPU, but neither of them reached a reasonable token speed.

We used Vercel's serverless function to build our backend. It mainly acted as a proxy to re-route users' requests to the real API address. We used Upstash for Redis, a serverless database to store users' credentials and validate their requests. The backend has two endpoints. One is for text responses and tool use cases, and another one is for image understanding. Any LLM requests will be eventually routed to Google's Gemini Flash 2.5 API to understand users' questions and execute tools if applicable. And if the "describe photo" tool is triggered during this, another request will be routed to our self-hosted VLM API address to caption the photo with user's instructions.

3.2 Evaluation

3.2.1 Usability Inspection. We did our own usability inspection and hands-on testing to check the frontend interface's accessibility. Before other tests, we usually did a simple hands-on testing using browser's "inspect" feature. We then used an Android 10 phone with TalkBack enabled to test how it works with a screen reader. We used WAVE web accessibility evaluation tool on the desktop to check any issues. Additionally, we added an automatic workflow to run a general accessibility test using Pa11y against the frontend website each time it is deployed.

3.2.2 Usability Testing. We attempted to recruit blind or low vision participants for usability testing through convenience sampling. We posted on Reddit with our study introduction and a screening survey link. We received four valid responses, including two from individuals who identified as low vision and two as blind. However, due to scheduling conflicts and time zone differences, we were only able to conduct usability testing with one participant who self identified as low vision. The session was conducted remotely via Zoom meeting. Participant's profile is shown in Table 1.

Prior to the session, we sent the consent form to the participant. At the beginning of the session, we verbally explained the consent form and obtained their agreement to participate. We then asked them to complete a series of tasks while thinking aloud. These tasks were: (1) Take a selfie using the system with any method, (2) view the photo after capturing, (3) use Voice Control Mode to take a selfie, (4) use simple Guidance mode to take a selfie, and (5) check if their face is centered or not in the camera view. After the tasks, we briefly asked the participant if they had used other tools to assist them in taking selfies and how they would compare those with our system.

4 Results

4.0.1 Usability Inspection Results. Due to fast iterations of our system, we did encountered many issues but can't list all of them here. There was an accessibility issue we discovered only after we tested with a screen reader on a smartphone. It was related to the switch button to switch between Voice Control mode and Guidance mode. We initially added a simple ARIA label to indicate it's purpose and updated its ARIA pressed status(on/off). But that's not enough for a screen reader user to know both the current mode and the next mode they are going to. To address this, we updated the ARIA label to include such information. An example of the label could be "Guide mode, switch to voice mode".

We also discovered that different browsers may have different implementations on certain APIs. For example, Edge and Chrome handle speech-to-text results very differently. Chrome streams the results while Edge just returns the final result. This caused us trouble since the system worked well when tested on Edge but malfunctioned on Chrome. We also found that our VAD module had a race condition with the speech-to-text module, which only happened when we tested on our Android phone, but not on a desktop browser.

4.0.2 Usability Testing Results. We identified several usability and accessibility issues during our usability testing. The most serious one was that there was no audio feedback to confirm something is completed. In task (1) the participant found the button to take the photo in a short time, but noticed there was no audio feedback to confirm the photo was taken. This also happened when the participant used their voice to asked the system to take a photo for them. For the first case, we had indeed forgotten to add feedback. However, we did implement feedback for voice commands. It turned out that the participant's TalkBack actively interrupted the verbal feedback. This issue did not occur on the Android 10 phone we used for internal testing. For the same reason, the audio guidance in the Guidance mode was also interrupted. The participant could only hear the guidance prompts like "move phone away" after turning off the screen reader.

In task (5), the participant used the guidance mode to check if their face was centered. We noticed that they did not use the "describe photo" feature. To use that, users could click on the camera view to receive an audio description even without using voice commands. When asked how we could make this feature more visible, the participant suggested adding an explicit button especially for screen reader users.

When asked about other assistive tools for taking selfies, the participant mentioned they were using a Pixel phone from Google with their Guided Frame feature. They thought the main difference between Guided Frame and our system is that Guided Frame will automatically capture the selfie for them, but our system requires manual capture. When asked which way they preferred, they replied that it depends. In some cases, such as taking selfies with friends, they prefer manual control. They also found it annoying when photos are accidentally captured while taking a landscape photo if someone walks in (although this is not related to selfies). They suggested adding a toggle to let users choose between automatic and manual capture.

We also found that the participant tended to use command-like phrases (like "take a photo") rather than natural language when using Voice Control mode in task (3). They also did not attempt to talk to the system when asked to check if their face was centered in task (5), instead relying on the audio guidance in Guidance mode. This suggests that the system's support for conversational interaction may not be discoverable enough to users. An initial tutorial could help new users understand the system's features.

As for what worked well, the participant took pretty short time to complete tasks using their preferred methods. Sometimes they had already finished one before we assigned it to them. They found the interface elements easy to locate, such as the capture button or the album navigation button, as they were all labeled and could be read out by the screen reader. They also found that the Simple Guidance mode required little time to learn due to its similarity to existing tools like Guided Frame. This suggests that our decision to include interactions familiar to users was a valid design choice.

5 Discussions & Reflections

The development of this web-based guided selfie assistant not only demonstrated the promise of designing conversational, accessible selfie tools for people with visual impairments but also the challenges. One of the major strengths of the system was how it drew upon familiar camera interface patterns, enabling this participant to accomplish tasks in a quick and confident manner. On the flip side, this level of familiarity restrained expectations the participant fell back to command-like speech and didn't experiment with the conversational capabilities of the system. That suggests that flexible, human-like interaction needs explicit introduction-through onboarding or examples-so that users can benefit from rich, descriptive guidance rather than treating the system as a simple command executor.

The project also served to reinforce the challenge of providing consistent audio feedback where screen readers are present. With TalkBack on, frequent interruptions to system speech reduced both Simple Guidance and Voice Control modes. This points towards one of the growing tensions for design in audio-first interfaces going ahead: future iterations should integrate subtle haptics to reduce the dependency on continuous spoken output.

Another insight concerns the user's reliance on real-time framing guidance over the interpretation that could be done post-capture. Even when natural language analysis was available, this participant always used the Simple Guidance mode when checking alignment-supporting the broader finding that BLV users must have real-time correction to prevent unfixable errors. Thus, this would suggest that the hybrid approach may best support users' workflow: concise directional cues with optional conversational detail.

The system's architecture was thus designed with considerations of privacy, keeping much of the image data local and depending on the backend minimally. This corresponds well with documented concerns by BLV users but will need better communication in subsequent versions to reassure users of data handling.

However, it remains a notable constraint that the participant pool was limited. While insights from the single low-vision session were extremely useful, further testing with blind and low-vision users, device types, and assistive technologies can help more deeply validate the robustness of the system. Taken together, this project indicates the feasibility of browser-based conversational selfie assistance in a way that articulates particularly how multimodal feedback, discoverability, and technical reliability need refinement for truly independent and expressive photography experiences.

6 Conclusion

We designed and implemented a web based selfie system for people with visual impairments. Users can ask question or express their needs to this system and receive natural language responses, similar to talking with a real human. Users can also choose to use a simple Guidance mode to only passively receive guidance on their position and distance. We conducted usability inspection during development and usability testing with one low vision participant. We found that clearly labeling web elements helped screen reader users adapt to our selfie system quickly. We also found it pretty helpful to let the user start with designs that they feel familiar. However, we discovered that without explicit prompts in a system, users may rely on their previous experiences and have difficulty discovering novel features, such as the ability to talk with the system using natural language. To improve this system, we need to conduct usability testings with more participants, since our sample size was limited to only one. We hope this work could provide insights for future development of accessible selfie systems, leading to a more independent selfie experience for people with visual impairments.

7 References

[1] Gonzalez Penuela, R. E., Vermette, P., Yan, Z., Zhang, C., Vertanen, K., & Azenkot, S. (2022). Understanding How People with Visual Impairments Take Selfies: Experiences and Challenges. Proceedings of the ACM on Human-Computer Interaction (ASSETS).

[2] Google Pixel Guided Frame. Official documentation and accessibility feature description (Google Support & Google AI Blog).

[3] Lim, J., Yoo, Y., Cho, H., & Choi, S. (2019). TouchPhoto: Enabling Independent Picture-Taking and Understanding for Visually-Impaired Users. Proceedings of the 2019 International Conference on Multimodal Interaction (ICMI).

[4] Yasmin, S. (2020). Haptic Selfies: Bold and Beautiful Living for the Blind and Visually Impaired. ACM Symposium on Virtual Reality Software and Technology (VRST).

[5] Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., & Bigham, J. (2018). VizWiz Grand Challenge: Answering Visual Questions from Blind People. CVPR.

[6] Bennett, C. L., Mott, M. E., Cutrell, E., & Morris, M. R. (2018). How Teens with Visual Impairments Take, Edit, and Share Photos on Social Media. Proceedings of the ACM on Human-Computer Interaction (CHI).

[7] Taslima Akter, Bryan Dosono, Tousif Ahmed, Apu Kapadia, and Bryan Semaan. (2020). "I am uncomfortable sharing what I can't see": privacy concerns of the visually impaired with camera based

assistive applications. In Proceedings of the 29th USENIX Conference on Security Symposium (SEC'20). USENIX Association, USA, Article 109, 1929–1948.

A Appendix

A.1 Source Code

Project repository: https://github.com/wyh2001/guided-selfie
Note: A user key is required to test the Voice Control mode in the deployed demo. Please contact the authors to request one if needed.