ImageExplorer Deployment: Understanding Text-Based and Touch-Based Image Exploration in the Wild

Andi Xu
University of Michigan
Ann Arbor, MI, USA
andixu@umich.edu

Minyu Cai
University of Michigan
Ann Arbor, MI, USA
jerrycmy@umich.edu

Dier Hou
University of Michigan
Ann Arbor, MI, USA
deerhou@umich.edu

Ruei-Che Chang
University of Michigan
Ann Arbor, MI, USA
rueiche@umich.edu

Anhong Guo
University of Michigan
Ann Arbor, MI, USA
anhong@umich.edu

ABSTRACT

Blind and visually-impaired (BVI) users often rely on alt-texts to understand images. AI-generated alt-texts can be scalable and efficient but may lack details and are prone to errors. Multi-layered touch interfaces, on the other hand, can provide rich details and spatial information, but may take longer to explore and cause higher mental load. To understand how BVI users leverage these two methods, we deployed ImageExplorer, an iOS app on the Apple App Store that provides multi-layered image information via both text-based and touch-based interfaces with customizable levels of granularity. Across 12 months, 371 users uploaded 651 images and explored 694 times. Their activities were logged to help us understand how BVI users consume image captions in the wild. This work informs a holistic understanding of BVI users' image exploration behavior and influential factors. We provide design implications for future models of image captioning and visual access tools.

CCS CONCEPTS

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

KEYWORDS

Image Caption, Alt Text, Touch Exploration, Screen Reader, Accessibility, Blind, Visual Impairment, Deployment, ImageExplorer

ACM Reference Format:
Andi Xu, Minyu Cai, Dier Hou, Ruei-Che Chang, and Anhong Guo. 2024. ImageExplorer Deployment: Understanding Text-Based and Touch-Based Image Exploration in the Wild. In *24th International Web for All Conference (W4A '24), May 13-14, 2024, Singapore.* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

W4A '24, May 13-14, 2024, Singapore

ACM ISBN 978-1-4503-XXXX-X/XX/XX…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Individuals who are blind or visually impaired (BVI) heavily rely on alt-texts to access visual content in images.[1] Creating these alt-texts demands substantial effort and time, resulting in most online images lacking accessible alt-texts [8] (e.g., over 60% images on websites lack them). In response, various tools and methods have been developed to make image captions more available, including crowdsourcing [7], web crawling [16], AI-driven techniques [22, 23, 37], and human-AI hybrid approaches [13]. Despite these advancements in promoting the caption authoring process, there is still little known about the consumption of image captions by BVI people, who have their own unique needs and goals on image captions [33].

The primary approaches to consuming image captions include *text-based* and *touch-based* exploration. Text-based exploration presents information sequentially in a list, which is generic and familiar for BVI users to use a screen reader to navigate them (Figure 1b). Touch-based exploration, on the other hand, conveys richer information spatially (e.g., relative positions, objects' sizes, details) to BVI users and enables them to perform touch gestures on the touchscreen (Figure 1c); while it leads to a higher cognitive load and a lengthier exploration time, and may cause users to overlook certain objects in an image [25, 29]. However, these insights were obtained through lab-controlled studies [25, 29]. Such settings may not fully capture the complexities and natural user interactions encountered in everyday life. There exists a knowledge gap in how BVI users interact with and weigh against the two methods when both are presented to users simultaneously in the wild.

In response, we deployed ImageExplorer, a free, publicly available iOS app on the Apple App Store built based on state-of-the-art approaches from Lee et al. [25] (https://imageexplorer.org/). Unlike iOS's built-in functionality for image description [4] and Seeing AI [28], which only provides object information for touch-based exploration in a single layer, ImageExplorer [25] enables multi-layer progressive information for touch-based exploration. We also added a text-based exploration interface to ImageExplorer, which allowed us to examine how end users access information via the two methods. In this paper, we aim to use the data we collected via the app deployment to answer the research question: **How do BVI users consume image captions of different modalities in**

[1] We use captions, alt-texts, and descriptions interchangeably in this paper.

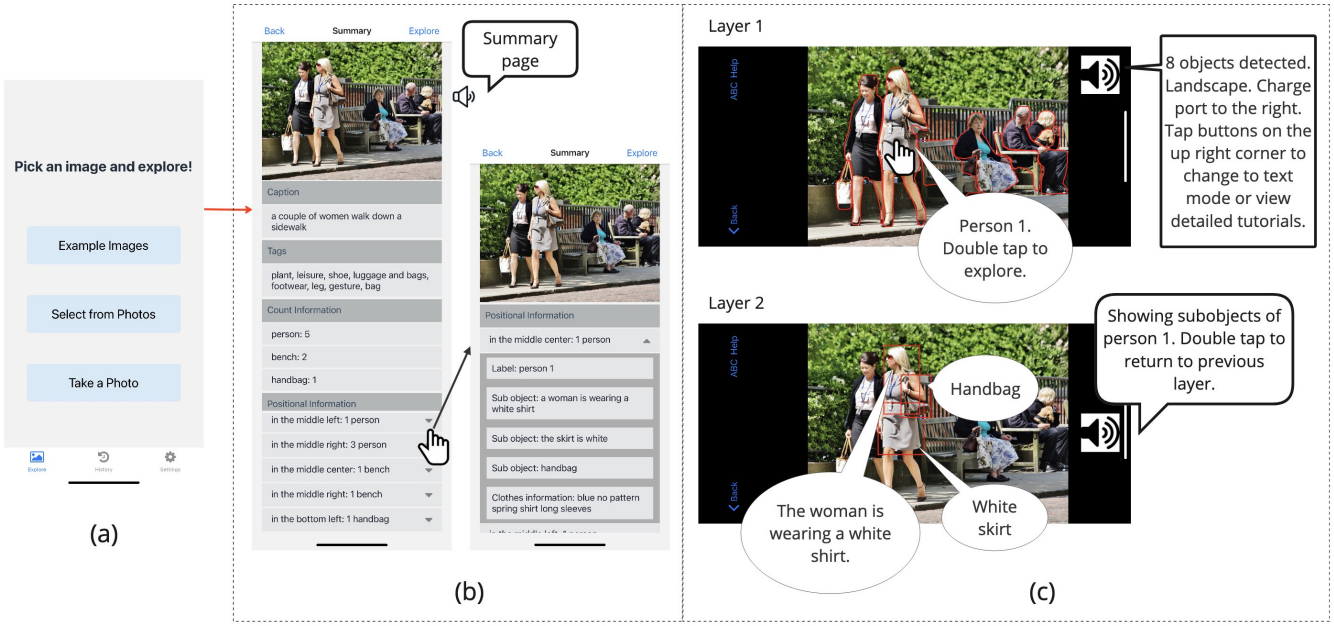


Figure 1: ImageExplorer User Interface. (a) The user can choose images from different sources, including example images we provide, their photo gallery, or take a new photo. After selecting an image, the app will automatically direct to (b) a text-based interface with summaries and expandable lists of detailed information, including their position on the image and object details. The user can click the "Explore" button to enter the touch-based interface, where (c) the user can perform touch gestures on the touchscreen to explore the image content with corresponding audio feedback.

the wild? We aim to understand the nuances of user interaction with these modalities in uncontrolled environments, addressing overlooked aspects in lab settings including multi-modal information consumption and the real-life challenges BVI users face. More specifically, we investigate:

(1) **What** are the primary categories of images users explored with different modalities?
(2) **How** do users use touch-based exploration?
(3) **Why** some users retain app usage while others do not?

Our research was approved by our institution's IRB. From March 1st, 2023 to February 10th, 2024, we collected 651 images, 487 of which were from 97 users who opted-in to our IRB-approved study and provided their accessibility settings. We analyzed the 487 opt-in images, and their corresponding 507 explorations (as one image can be explored multiple times). We found that BVI people were most interested in accessing images featuring persons and objects. They were more likely to use touch-based explorations for images featuring people, document and setting, while accessing images with objects via text-based exploration.

Furthermore, the number of primary targets in an image also related to the users' preference for text-based and touch-based exploration. We found that users were more likely to use touch exploration when there is more than one primary target in the image. Moreover, during touch exploration, users tended to start around the center area (Figure 8), moved their fingers over the images, and slowed down their movement when touching an object. Also, when they encountered inaccurate captions, users could notice the inaccuracies and retake another better photo. Lastly, we found that the accuracy of captions and the number of objects users could discover in touch-based exploration influenced user retention.

In summary, we studied how BVI users used text-based and touch-based methods to access images in the wild, and identified their preferences and strategies in the two modes. We further discuss our lessons learned and provide implications for future research on designing accessible image captions and explorations.

2 RELATED WORK

Our work on deploying ImageExplorer builds upon image accessibility, tools to improve image captions availability, and prior studies on system deployment for BVI people.

2.1 Image Accessibility Issues

BVI people need captions for images from public content, such as the web. The importance of providing descriptions for images to ensure web accessibility has been emphasized by the Web Content Accessibility Guidelines [10]. However, a significant portion of images on prominent websites still lacks descriptions, rendering them inaccessible to BVI users [8, 31]. The emergence and dominance of social media platforms, teeming with user-generated content, further exacerbate this issue. A mere 0.1% of over a million tweets with images were accompanied by captions, with users either forgetting or not knowing what to include in these captions [12]. Additionally, there are also needs from BVI people to understand image content

generated by themselves, like photos taken by BVI people [6]. Capturing well-framed photos is challenging for BVI people because they cannot get visual feedback of photos they took. Prior work also explored how to provide the user with sound, vibration or other feedback to help BVI people get feedback when they are taking photos [2, 5, 20, 21, 34–36]. The issue of such two sources of images accessibility issues have thereby motivated researchers to develop efficient tools and methods to make images more accessible.

2.2 AI-powered Tools to Improve Image Caption Availability

Given the evident gaps in authoring image descriptions, several methodologies have been proposed, which range from human-based crowdsourcing [7], machine-based web crawling [16], AI-driven techniques [22, 23, 37], or human-AI hybrid approaches [13]. Recent new, more powerful AI models make it possible for several platforms to capitalize on these advancements. For instance, Facebook's Automatic Alt Text system [3], Google Chrome's image description functionality [15], and Seeing AI [28] all offer comprehensive image descriptions using computer vision models. After the release of GPT-4V [30], BeMyEyes released a new feature, BeMyAI, which lets users chat with an AI assistant to get detailed descriptions on images [11]. Although the development of tools to assist in image caption creation is an active area within the BVI community, the understanding of how BVI individuals can more effectively utilize these captions — considering their diverse needs in real-world scenarios — remains under-explored. In our work, we thus investigate how BVI users consume image captions in the wild.

2.3 Text-based and Touch-based Image Exploration

Beyond text-based exploration, Seeing AI [28] enables BVI users to explore objects and their captions by touch. In contrast to Seeing AI [28], which only has one layer of information, Lee et al. [25] provides hierarchically layered captions on objects that entail progressive details and spatial information; its rich information also helps users better identify wrong captions generated by AI, a challenge identified in [27]. However, with touch-based exploration, users might encounter difficulties in locating all the detected objects within an image, and accessing all information can be more time-consuming than a text-based counterpart. To improve the touch-based experience, ImageAssist [29] introduces a combination of tools, such as a menu, beacon, and hint tool, to offer an image overview and facilitate the identification of key image areas. Considering these works, we built our own ImageExplorer system by integrating a text-based summary and touch-based functions with hints and customizable settings, both of which provide general image captions and layered object details. With the deployment of ImageExplorer on the Apple App Store, we are interested in understanding how users perceive text-based and touch-based exploration in the wild.

2.4 Deployment Studies with BVI People

Several systems for BVI users have shifted from research to deployed applications, which can provide valuable large-scale real-world data to help people understand BVI people's needs. The deployment study of VizWiz [9], for instance, reported features of questions asked by BVI users and subjects they are interested in images, which later helped researchers better understand BVI users' needs for image accessibility [17, 18]. Microsoft's Soundscape [1], on the other hand, leverages 3D audio technology to foster richer environmental awareness and navigation, and its deployment data [26] contributes to a more holistic understanding of important features associated with user retention and app usage. Gonzalez et al. conducted a two-week diary study with 16 BVI participants to study how they use an AI-powered scene description application [14]. Our present study focuses on a large-scale deployment to understand user interaction patterns with both text-based and touch-based image exploration, using realistic logs in the wild.

3 DEPLOYING IMAGEEXPLORER

ImageExplorer provides BVI users with text-based descriptions (Figure 1b) and multi-layer touch interfaces (Figure 1c) to help them understand image content. The method we used to generate layers for touch exploration is based on the state-of-the-art approach by Lee et al. [25]. We deployed our ImageExplorer's back-end server to the Google Cloud Platform and communicated all the data using Firebase. We support more functions and improve usability in our deployed version as follows. A demo video is available at https://youtu.be/fQ-QPNvGAT4.

Once the user uploads an image (Figure 1a), the server processes the image through a series of AI models and structures all information into a multi-layer tree, which takes about 10-15 seconds (Figure 2). ImageExplorer then organizes the information tree into a text-based summary page (Figure 1b) and a touch-based multi-layer interface (Figure 1c). The text-based interface is similar to Facebook's Automatic Alt Text system [3], which contains a caption, tags related to the image, and count of each detected objects. To also increase spatial awareness of text-based exploration, we offer location details (e.g., middle left, top left) on the text summary page, where the location information is stored in an expandable list with captions of object details in a hierarchical manner.

Entering the touch-based interface, ImageExplorer announces the total number of objects, the image orientation, and first-layer elements. When users move their fingers on the image and do not touch any objects, background music plays; when an object is touched, its name will be read out (e.g., person, bench); and if the touched object contains sub-objects or details, audio feedback will prompt the user *"double tap to explore"* at the end of the caption. In the subsequent layers, the system displays corresponding elements, and users can go back to the previous layer by double-tapping anywhere else on the screen. Users can also toggle the button of text recognition mode to recognize text in the image.

Additionally, ImageExplorer has a *History* page (Figure 3a) for users to access all of their previously uploaded images; and a *Settings* page (Figure 3b) for users to customize specific information they prefer (e.g., clothes, pose, and pet), and the granularity of the results,

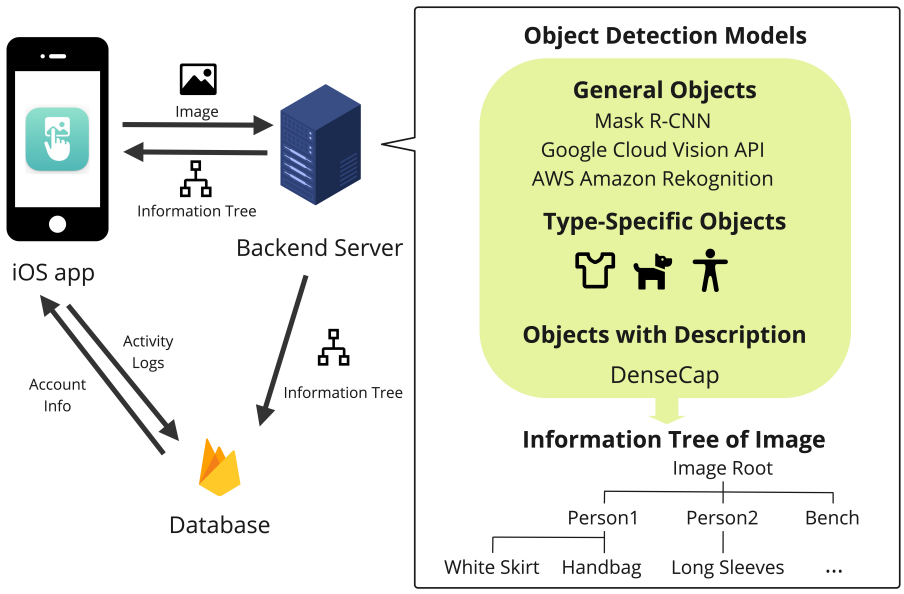
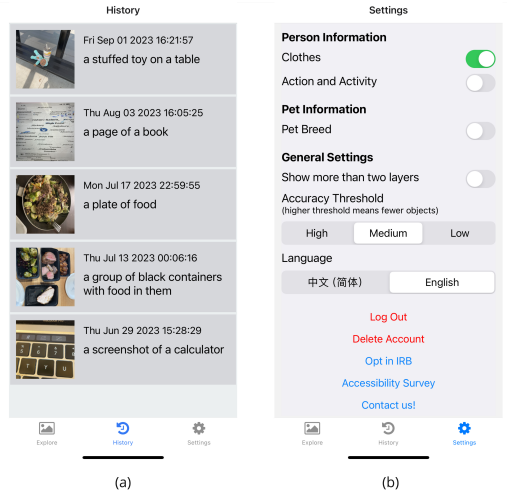


Figure 2: ImageExplorer System Diagram. The app sends images to the backend. The backend will process the image through a series of models and return a hierarchical information tree of the image. Firebase is used to store all relevant data.

such as the number of layers of object details, or the accuracy threshold of models. The system is bilingual in English and Chinese.

4 METHODS

We collected real-world usage data from our ImageExplorer app, and performed both quantitative and qualitative analysis to understand BVI users' consumption of different modalities of image captions.



(a) (b)

Figure 3: ImageExplorer *History* (a) and *Settings* page (b). The *History* page displays all previous explorations, and users can adjust their preference for different types of information in the *Settings* page.

4.1 Data Collection

Upon installing the app, users were prompted to read how we will collect and use their data and be asked whether to opt into our research with informed consent provided. User account data associated with personal information were stored separately from research data. All other data was tagged with anonymous IDs. The data we collected for our study falls into five categories:

(1) **Accessibility settings**: We set a survey in the app to collect users' accessibility settings. Options include VoiceOver, Magnifier, Spoken Content, Dictation, Text Size, Others, and None.
(2) **Customization settings**: This includes language settings, information needs (e.g., poses, cloths, pet breed), and granularity of the results, such as the accuracy threshold of models and the number of layers.
(3) **App activity logs**: This captures actions including sign-in, image uploading, and other activities on navigating functions such as buttons pressed, mode switching, etc.
(4) **Text-based activity logs**: Textual items that users look into in the text summary page.
(5) **Touch-based activity logs**: Touch data, including the exploration trajectory, gestures (e.g., single, double tap), and layers that users delve into.

From March 1, 2023, to Feb 10, 2024, ImageExplorer had a total of 371 users, among whom 231 users uploaded and explored 651 images (Figure 4). Among the 231 users, 188 opted in to our IRB-approved study and contributed 566 out of 651 images for our analysis. Among the collection of 566 IRB-opt-in images and 6 example images we provided, users revisited and explored 725 times in total. In terms of accessibility settings, we collected 97 valid responses, as this survey is optional. Among the responses,

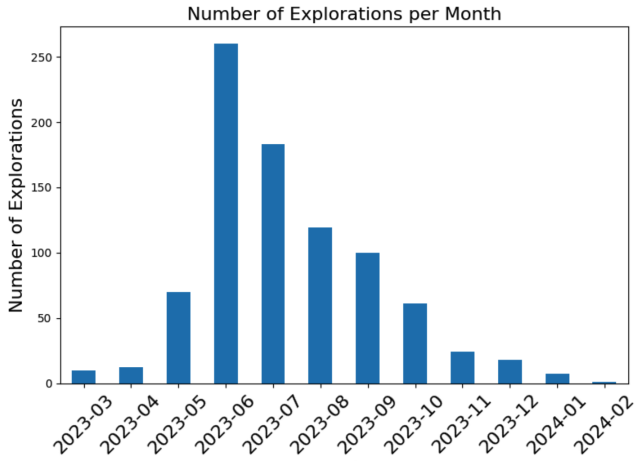


Figure 4: The growth of new explorations from March 2023 to February 2024.

77 users selected *VoiceOver*, *Spoken Content*, or *Dictation* (which could indicate fully blind), 19 users users chose *Magnifier* or *Text Size* (which could indicate low vision), and 1 user selected *Other*. Those 97 users generated 507 explorations on 487 self-uploaded images and 91 explorations on 6 example images provided by the app. Therefore, the deployment data reported is based on the 507 explorations from the 97 users.

Additionally, during the deployment, we sent out optional surveys to users to ask about their experiences and suggestions on the app and received 4 responses. Our survey collected information about the user's usage context and strategies when using ImageExplorer, including questions about photo sources, motivation to use the app, strategies when using the two modalities, and etc. Since the survey was optional and we only received 4 responses, we only use this data to supplement our log data.

4.2 Data Analysis: Grouping Based on Exploration Modalities

To understand how users consume different modalities of image captions, we categorized 487 user-uploaded images into *Touch-Explored* and *Text-Explored* groups. Among the 507 explorations logged, users either explored images through text mode only or with both text and touch mode. These results stem from the app's inherent design: the app automatically directs to the text-based page after processing the uploaded image. Consequently, a user cannot engage solely in touch exploration without first encountering the text summary. A user can either only use text in one exploration, or use both text and touch in one exploration. Therefore, we categorized images into two types based on the exploration modalities they have been used with:

(1) ***Text-Explored* images**: images which are only explored with text exploration (N=274).
(2) ***Touch-Explored* images**: images which are explored with touch exploration (N=233).

4.3 Data Analysis: Grouping based on Image Content

We categorized images based on their primary categories to understand image content from our users, to learn whether it influenced users' consumption of different modalities. We used the same classification schema presented by VizWiz [9] and additionally added *Document* to describe screenshots or photos of text-rich documents and posters. The resulted labels are **Primary Object Labels**: `Object`, `Setting`, `Person`, `Animal`, `Document`, `Error`, and `Unclear`. `Error` refers to images with blurry content or are too dark to be read, and `Unclear` refers to images with good quality, but is hard to find primary subjects.

Additionally, we investigated how these images were captured regarding clarity and focus. We classified images using the following **Target Number Labels**:

(1) *Single-Target*: featuring one clearly visible object as the focal point;
(2) *Multi-Target*: containing several visible objects, making it challenging to discern a primary focus; and
(3) *No-Target*: either too blurry or overcrowded, so no particular object stands out as the primary target.

These three categories stemmed from the process where two researchers randomly picked 42 images (10%) and separately labeled them, compared their labels to resolve conflicts and achieved a consensus on the rubric for categorization. The two researchers then labeled another 42 images based on the rubric and achieved Krippendorff's Alpha-Reliability [24] of $\alpha = 0.96$. After that, each of the two researchers labeled half of the remaining images.

4.4 Data Analysis: Accuracy of Machine Generated Captions

Lee et al. examined touch interface's ability to encourage users' skepticism on wrong captions, and showed that touch interfaces can work better in this scenario [25]. We wanted to see whether this finding is consistent in our deployment study. Besides the activity log data, we manually examined the machine-generated labels of the images to assess their accuracy. In this context, an "inaccurate" label denotes false positive captions (when a label refers to something not present in the image). For example, in one image, a man is holding his hand in a fist, but the caption says he is holding a knife. It is worth noting that our accuracy analysis did not encompass instances where the machine omitted any objects, so we ignored false negative captions. Some captions included attributes to objects, like "blue clothing." We did not take the correctness of attributes "blue" into account as long as "clothing" is a right caption. To further delineate the nature of these inaccuracies, we categorized them as either "first-layer inaccuracies" (where the label inaccuracy occurs in the first layer) or "sub-layer inaccuracies" (where the label inaccuracy is located in a sub-layer).

5 RESULTS

In order to answer the research question, **how do BVI users consume image captions of different modalities in the wild**, we break it done into three sub-questions: **what** are the primary categories of images users explored with different modalities; **how**

do users use touch-based exploration; and **why** some users retain app usage while others do not. We analyzed the usage log data on the 487 user-uploaded images from our app, and presented the answers to those three questions in this section. We present all results in terms of percentages, and acknowledge these results should be interpreted with caution without statistical tests to determine significance.

5.1 What are the primary categories of images users explored with different modalities?

To investigate this question, we analyzed the association between images' exploration modalities (either *Touch-* or *Text-Explored*) and their primary categories of content (Table 1).

5.1.1 Distribution of Primary Object Labels. In the 487 images (excluding app's 6 example images), for **Primary Object Labels**, most of the explorations (27.7%) are `Person` images, 24.5% for `Object`, 14.6% for `Setting`, 13.4% for `Document`, 12.6% for `Unclear`, 3.4% for `Animal`, and 2.4% for `Error` (Figure 5).

We observed user preference patterns for image exploration modalities based on **Primary Object Labels**. Specifically, 51.0% `Person` is *Touch-Explored* and 49.0% is *Text-Explored*. Conversely, 44.4% of `Object` is *Touch-Explored* while 55.6% is *Text-Explored*. `Setting` has 51.4% in *Touch-Explored* and 48.6% in *Text-Explored*. Regarding `Document`, 58.0% is *Touch-Explored* and 42.0% is *Text-Explored*.

The larger ratio in *Touch-Explored* group might suggest that users tended to use a touch-based method to explore images related to `Person`, `Document` and `Setting`, while this was the opposite for `Object`. As for `Setting` images, a plausible explanation for this is their intrinsic complexity which encompasses numerous objects, such as an array of furniture in a room. None of the `Setting` images were `Single-Target`. Users needed to grasp the interrelations between those objects, especially their spatial layout, and touch-based exploration can be a more helpful modality in such cases.

5.1.2 Distribution of Target Number Labels. For **Target Number Labels**, 57.0% of `Single-Target` is *Text-Explored*, while 43.0% is *Touch-Explored* images. 47.6% of `Multi-Target` is *Text-Explored*, while 52.4% is *Touch-Explored*. 54.6% of `No-Target` is *Text-Explored*, while 45.4% is *Touch-Explored*.

Category (Frequency %)	Text (%)	Touch (%)
Person (27.7%)	49.0%	51.0%
Object (24.5%)	55.6%	44.4%
Setting (14.6%)	48.6%	51.4%
Document (13.4%)	42.6%	57.4%
Unclear (12.6%)	70.3%	29.7%
Animal (3.4%)	76.5%	23.5%
Error (2.4%)	66.7%	33.3%
Single target (40.8%)	57.0%	43.0%
Multi target (38.3%)	47.6%	52.4%
No target (20.6%)	54.6%	45.4%

Table 1: Percentage distribution of text and touch modes, ordered by category frequency.

Single-target images have a focused subject, and from our observation, textual descriptions can already efficiently convey the necessary information about the subject. In contrast, in images with multiple targets, besides the details of each object, the spatial dynamics and relative positions of these objects also become crucial for comprehensive understanding. Touch-based exploration can satisfy this need, allowing users to discover the spatial relationships between different targets. Thus, in multi-target scenarios, touch provides a richer, more nuanced understanding, making it the preferred modality.

5.2 How do users use touch-based exploration?

Next, we used the app activity logs defined in section 4.1 for this analysis, and break it down into the following.

5.2.1 Users' Finger Moving Pattern. Figure 6 shows the amount of time users spent on touch explorations, and 86.7% users spent less than 150 seconds exploring an image. Specifically, we found that users tended to slow down and linger over specific areas when they received audio feedback. If the object did not receive the user's interest, subsequent passes over the same area were marked by a swifter finger movement. Conversely, if the object captured their curiosity, they frequently revisited the area, touching it repetitively and listening intently to the audio feedback, with their fingers moving at a more deliberate pace. We also noticed that users moved their finger following a circular (Figure 7a) or zigzag pattern (Figure 7b) when exploring. The 4 responses from our survey aligned with our observation. Two users said they usually slow down when using touch exploration when finding objects of their interests. Two users said they would slow down when finding audio output unexpected or wrong. One user said that they just generally slow down when listening to the application's audio description. In another question of how they touch and move their fingers, 2 users said they move fingers in a zigzag way, and 1 user in a circular way, to explore objects in images. For each touch exploration, we picked out their fingers' starting position. These positions were normalized within the range [0,1], and a heatmap (Figure 8) was generated to visualize the distribution of all starting points. Notably, a concentration of starting points was observed in the central region of the image. This observation may indicate a user tendency to start exploration from the middle region, likely because they anticipate that the primary object depicted in the image is located in the center.

5.2.2 Users' Responses to Inaccurate Captions. 431 explorations had no inaccuracies in the captions, and 38.7% of them were explored with touch. In contrast, among the 147 explorations with inaccurate captions, 44.9% were explored with touch. However, 44.9% is still not very high, so we cannot claim that inaccurate captions prompt users to use touch exploration more. A possible reason why users still stick to text exploration when seeing skeptical captions is that the current system also provides detailed information on the text summary page, providing users enough details from the text mode alone. Another common reactions we noticed from the dataset is that when users received an inaccurate caption from the app, they would retake another photo to try whether the caption could improve, and a lot of the times the retaken images had higher quality in terms of camera aiming angles and position of targets,

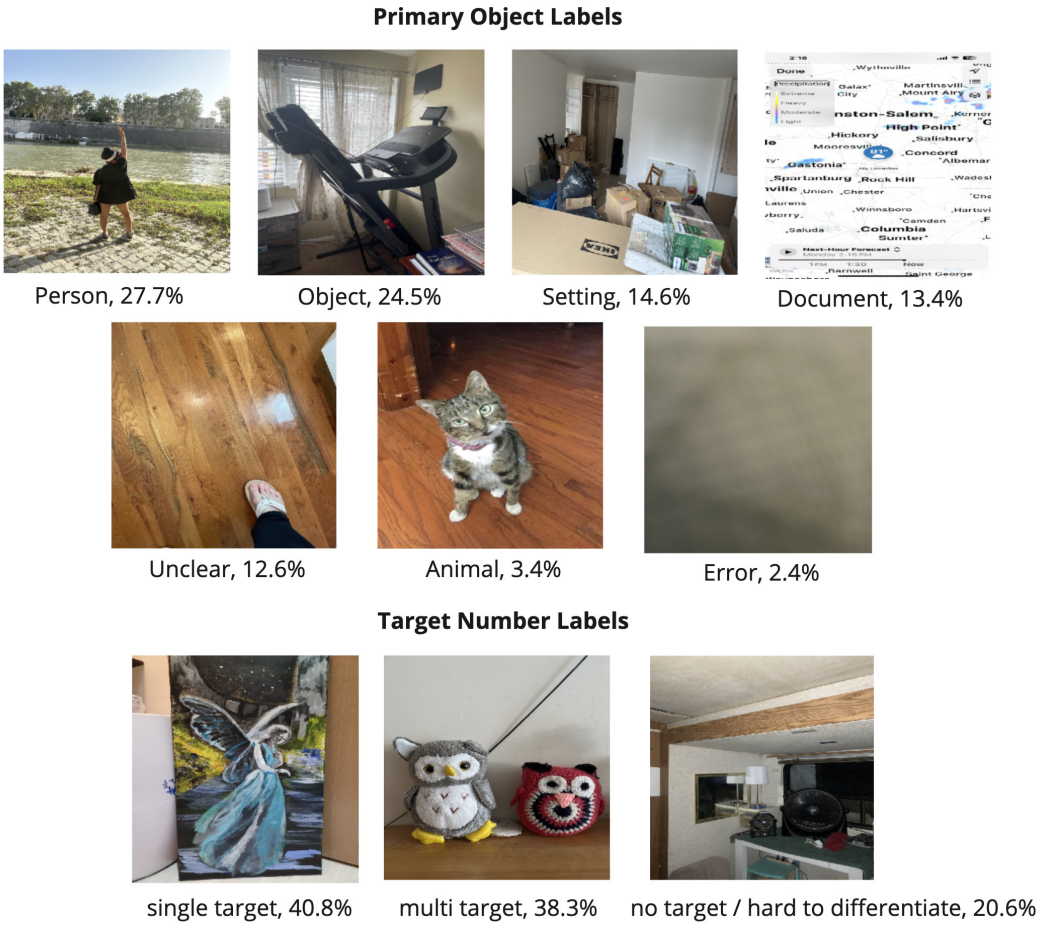


Figure 5: The Taxonomy of Image Categories. Each category is presented with a representative image sampled from the 487 images.

like the two images in Figure 9. The user first took image 9a, whose captions are "cup, home appliance, a black and white bag, the top of the fire hydrant." The user looked at the text summary page only. They then took image 9b, which had captions "toilet, bottle, boxed packaged goods, table, the box is white, the counter is white, the wall is white, the water is silver, a white cord, a white toilet paper, the bag is white, a black wheel, the white cord." The user used touch exploration this time. Even though the second round of captions still had errors, it provided more information because image 9b had better camera aiming and target positioning. This might show that users suspected the inaccurate captions and did something to resolve it.

5.2.3 Object Discovery Rate. Another problem with touch interfaces is a higher cognitive load for users to navigate and effectively find all detected objects in the image [25, 29]. To examine whether incorporating a text-based summary page could alleviate this problem, we quantified the discovery rate of first-layer objects using the following formula:

$$\text{Discovery Rate (First-layer)} = \frac{\text{\#1st-layer Objects Discovered}}{\text{\#1st-layer Touchable Objects}}$$

We also calculated a more general discovery rate using:

$$\text{Discovery Rate (General)} = \frac{\text{\#All Object Discovered}}{\text{\#All Touchable Objects}}$$

We looked at the 233 explorations where users used the touch interface. For first layer objects, the mean discovery rate is 39.1% with a standard deviation of 46.7%. Median rate is 0.0% and the third quantile is 100%. For the general object discovery rate, mean is 30.3% with standard deviation of 37.4%. Median is 0.0%, and the third quantile is 62.5%. The lower discovery rate on the general side is probably due to higher difficulty in finding all second layer objects, which are unavoidably smaller. If users missed some objects in the first layer, they would also miss its corresponding second layer details.

5.3 Why some users retain app usage while others do not?

Lastly, to understand user retention, we compared the activity logs of first-time-dropout users and active users. We define first-time-dropout users as individuals who only explored a single image and refrained from further app usage. Active users are users who

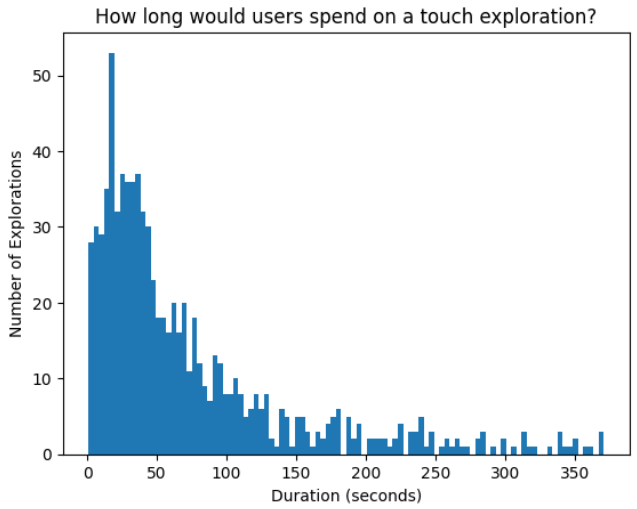


Figure 6: Time users spent on touch exploration. 86.7% users spent less than 150 seconds exploring an image.

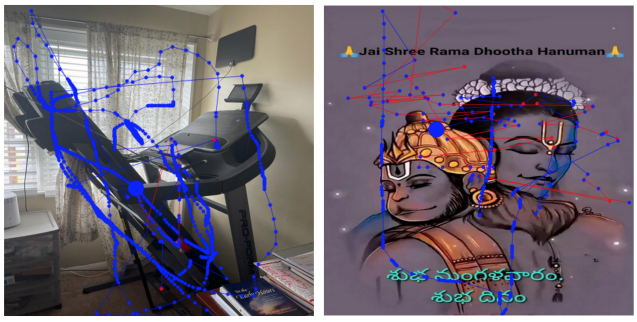


Figure 7: Two images uploaded by users, overlaid by their finger moving trajectories. Points are where they stopped, and lines show the trajectories. There is some circular movement in the background area in (a), and zigzag patterns in (b).

explored more than 4 times in our app. There are 37 first-time-dropout users and 28 active users.

5.3.1 Quality of Labels. Looking at the accuracy of image labels, a slightly higher proportion (37.8%) of images uploaded by first-time-dropout users contained inaccuracies in the generated labels. This percentage was slightly lower for active users, at 28.4%. This might suggest that the accuracy of initial labels could play a role in influencing a user's decision to continue using the app. A deeper examination into the inaccuracy rates across different layers of information reveals more granular insights. While the inaccuracy rate for the primary layer of information was marginally higher for first-time-dropout users at 18.9% compared to active users at 15.1%, a larger difference was observed in the second layer of information. First-time-dropout users experienced a higher inaccuracy rate of 35.5%, compared to the rate of 22% for active users. This might indicate that inaccuracies in detailed, secondary information could be especially discouraging for new users. This observation makes sense because one of the features of ImageExplorer that differentiates it from other similar apps is its ability to provide multi-layer information.

5.3.2 Categories of Images Uploaded. The second most common category for first-time-dropout users was `Document` at 21.6%. The third place was occupied by the `Unclear` category at 16.2%. Both of these categories have a propensity for generating inaccurate captions, potentially contributing to worse user retention. On the other hand, active users predominantly uploaded images with the theme of `Object` and `Person`, constituting 49.3% of their uploads. These categories have been observed to have the highest rate of accurate caption generation. The `Unclear` category was less prevalent among active users, ranking third at only 13%, less than its proportion among first-time-dropout users. Other themes like `Setting` and `Document` accounted for 14.2% and 9.8% respectively for active users.

5.3.3 The Ability to Find Objects in Touch Mode. We calculated the discovery rate using the same way as in 5.2.3 for the two groups. For first-time-dropout users, the mean discovery rate for first-layer objects is 36.4%; for general-discovery rate, mean is 27.8%. For active users, first-layer discovery rate has mean = 45.1% and median = 8.3%. General discovery rate has mean 35.2% and median 21.5%. There is an obvious difference between the rates of the two groups (Figure 10).

5.4 Unique Use Cases

We also share 3 interesting use cases we discovered in our survey responses. One user stated her motivation for using ImageExplorer was to commemorate her husband who passed away; she explored the photos of them to get as much information as possible. Even though she *"got so much detail from exploring with app, more than using any other app"* that she have used, she still hoped to get more details regarding their facial expressions and postures, like how her

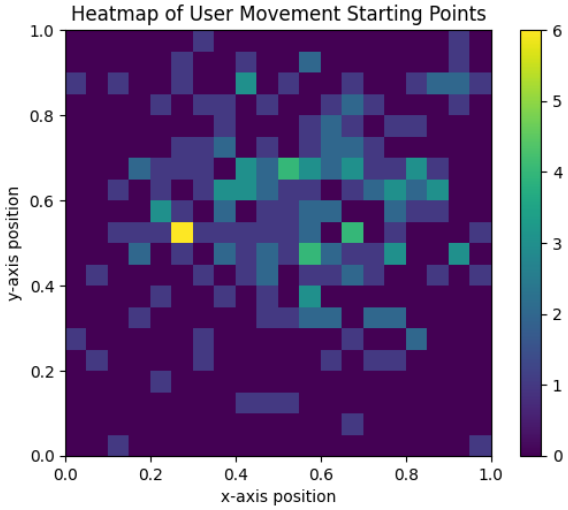


Figure 8: Heatmap of users' finger movement starting point, normalized to x= 1.0 and y=1.0 from different dimensions of user-uploaded images.

(a) (b)

Figure 9: The user first took image (a) then (b). (a) generated captions with errors. (b) has better camera aiming and target positioning.

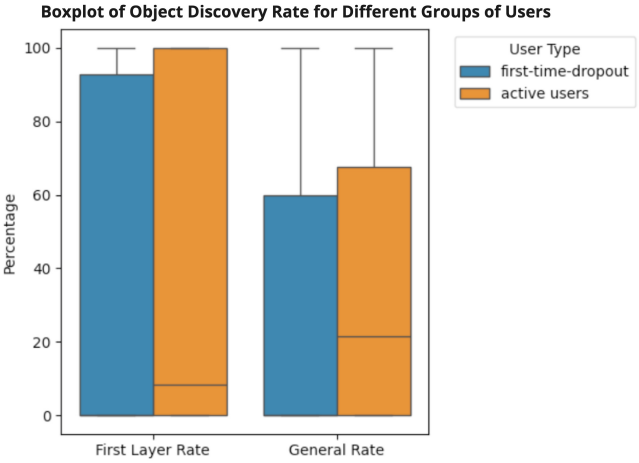
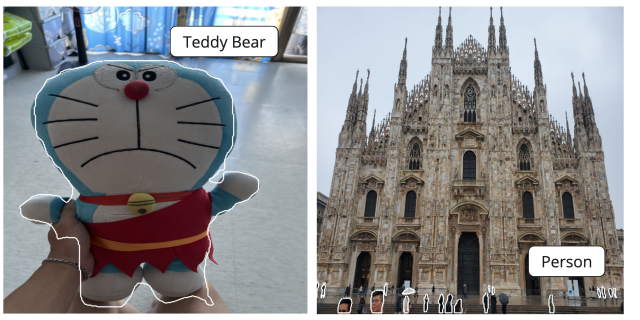


Figure 10: Boxplot comparison of object discovery rate between first time dropouts and active users.

husband was holding her with an arm. Two other users reported that their motivation to use touch exploration were to ensure they took good pictures and had the target objects in the middle of the image. One example was when one wanted to sell some of their used belongings. They said that *"if I'm taking a photo of something to sell or give away, it would be nice knowing if it is more or less centred."* Another was an author who wanted to take good pictures for their book. *"This is extremely handy for re-taking pictures of the scene – especially when I don't have the subject centered or elsewhere within the photo, or I've neglected to include something on the border."*

6 DISCUSSION AND FUTURE WORK

From the above results, we have observed multi-faceted interaction patterns in BVI people who consumed images in the wild. In this section, we discuss our lessons learned in the deployment and the implications of the results for future research on designing image captions and access tools.



(a) (b)

Figure 11: Two images uploaded by users that had inaccurate captions. (a) is a Doreamon, but the system detected it as a teddy bear. For (b), the building should be the focus, but the system only detected the people in front of it.

6.1 Implications for Designing Image-to-Caption Systems for End Users

With our analysis, we have pinpointed several critical challenges inherent to contemporary computer vision models, especially when deployed in real-world contexts:

(1) **Increasing Variability in Dataset Composition:** As we deployed ImageExplorer in a worldwide scale, certain categories of objects, especially those imbued with cultural differences, are susceptible to misidentification. For instance, a Doraemon doll was mistakenly detected as a teddy bear (Figure 11a), or pictures of culturally significant deities are often provided with inadequate descriptions (Figure 7b). Future work could explore how to provide richer datasets to accommodate different cultures to create adaptive image captions for different people.

(2) **Identifying the Focus:** We found that `Setting` images are naturally complex because of their rich elements. For example, a photo of a tourist spot might focus on people in the foreground rather than important landmarks such as monuments, buildings, etc (Figure 11b). These images may also suffer from the narrow field of view of the camera and the shooting perspectives, making it hard for models to fully capture and recognize the scene. Furthermore, when multiple objects are present, the model may fail to discern each object's individual identity or significance. Prior work by Stangl et al. [33] had identified that the context to access images is pertinent to the needs when consuming image captions. Thus, it is important to incorporate users' context and needs into the models.

6.2 Understanding Users' Context for Personalized Image Captions

We define users' context to be composed of the users' personal background, the story behind the image they uploaded, and their intention of exploring the image. We see an association between users' context, the exploration modalities they chose, and their desired image captions.

From results in section 5.1.1, we see a stronger preference to use touch-based method to explore `Document` type of images. This may be explained by the findings by Herskovitz et al. [19] regarding BVI people's challenges in existing applications: *these tools read extensive amounts of text at one time and do not allow users to pause or only listen to the information they need.* Therefore, users might lean towards touch exploration when navigating images containing text-based documents depending on their own needs.

When users use touch exploration, and pause or slow down in the process, based on the survey result, it could signal that they are interested in that region, or they deem the caption incorrect. In response to either intention, the system could run another prediction with a different model. This way, we could give the user different information to understand the content or help them identify the wrong captions by comparing different models' output [25]. Thus, by analyzing users' touch behavior, it is promising to provide in-situ and adaptive captions for end users.

Additionally, from the stories shared by users in section 5.4, we can see that behind each image, there is a unique story, and it is necessary to generate unique and personalized captions for individual users, similar to the findings in Stangl et al. [32] that the context of accessing images affects users' needs for image captions. Future HCI work could explore ways to provide personalized or potentially generate captions on the fly according to users' background, context, and usage behaviors on the app.

6.3 Conversational Image Exploration

In the scope of our study, we primarily focused on text-based and touch-based modalities. However, additional interactive modalities could also be helpful for BVI users when accessing images. One possible modality is conversational interfaces. Several applications have incorporated Large Language Models to facilitate the understanding of images. For instance, Be My Eyes has introduced Be My AI, a virtual assistant powered by GPT-4 [11, 30] which enables users to send images to the agent and receive content descriptions via dialogues. Such conversational interfaces allow users to delve deeper into specific information gradually; and the AI could provide more specific captions tailored to user's needs by incorporating their previous conversation context. This modality could augment touch and lower the cognitive load for BVI users via natural language. For instance, when hovering over the object of interest, the user can ask questions in situ to get more details, or they can ask the system to provide hints to help them find objects of interest, which could address the problem that users have a hard time in observing all provided information in touch-based exploration [25].

6.4 Limitation and Future Work

A limitation of our present study is the scale of the data collected. A larger, richer dataset could potentially allow for more robust and definitive conclusions. Additionally, all findings are presented in terms of percentages, and we acknowledge that without statistical tests to determine significance, these results should be interpreted with caution. The absence of statistical significance testing limits our ability to draw firm conclusions about the generalizability of these findings to the broader population. Future research could address this limitation with a larger user size and employing statistical methods to validate the observed trends. Finally, while we attempt to interpret the results derived from the qualitative data by deducing users' motivations, it is imperative to note that we lack a comprehensive understanding of users' underlying intentions and do not possess a definitive ground truth. "Active users" who used the app more than four times may contribute to a large portion of dataset and cause bias in our deduction of user behavior. It is also possible that fully blind users and low-vision users have different needs that cannot be captured by our analysis, due to the fact that we do not have full demographic data. Besides continuing to collect data, in the future, we plan to conduct other analyses and collect other types of qualitative data. It is helpful to gain an in-depth understanding of users' demographics, their motivations, the context in which images are sourced, and images' subsequent destinations. Comprehending the entire user journey, combined with users' exploration patterns, could offer insights into their desired information, aiding in the understanding of image caption consumption in the wild. A method to achieve this richer understanding would be to conduct diary studies and follow-up interviews with active users and ask about their use cases. Their direct feedback could provide valuable perspectives on the system's utility and areas of improvement.

7 CONCLUSION

In this paper, we reported how BVI people interacted with the iOS app ImageExplorer, an application that helps BVI people access image content via both text-based captions and multi-layer touch-based interfaces. By analyzing user-uploaded images and their usage logs across twelve months, we learned the factors influencing their preferences over text and touch modalities, image categories they were interested in accessing, their interactions with wrong captions, and factors influencing user retention. We hope our findings can provide insights for future AI research and assistive tool design.

ACKNOWLEDGMENTS

We sincerely thank all of our participants for their usage of the app, and their valuable feedback and suggestions. We thank the organizations and individuals who helped with spreading the word and study recruitment. We thank our lab members, including Jiani Huang and Hyeji Han for their help. This research is supported in part by a Google Cloud Platform Credit Award.

REFERENCES

[1] 2023. Microsoft Soundscape. https://www.microsoft.com/en-us/research/product/soundscape/

[2] Dragan Ahmetovic, Daisuke Sato, Uran Oh, Tatsuya Ishihara, Kris Kitani, and Chieko Asakawa. 2020. ReCog: Supporting Blind People in Recognizing Personal Objects. 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, 1–12. https://doi.org/10.1145/3313831.3376143

[3] Facebook Meta AI. 2021. Facebook Alt-text System. https://ai.facebook.com/blog/how-facebook-is-using-ai-to-improve-photo-descriptions-for-people-who-are-blind-or-visually-impaired/

[4] Apple. 2023. Use VoiceOver for images and videos on iPhone. https://support.apple.com/guide/iphone/use-voiceover-for-images-and-videos-iph37e6b3844/ios

[5] Jan Balata, Zdenek Mikovec, and Lukas Neoproud. 2015. BlindCamera: Central and Golden-ratio Composition for Blind Photographers. In *Proceedings of the*

Mulitimedia, Interaction, Design and Innnovation (Warsaw, Poland) *(MIDI '15)*. Association for Computing Machinery, New York, NY, USA, Article 8, 8 pages. https://doi.org/10.1145/2814464.2814472

[6] Cynthia L. Bennett, Jane E, Martez E. Mott, Edward Cutrell, and Meredith Ringel Morris. 2018. How Teens with Visual Impairments Take, Edit, and Share Photos on Social Media. 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, 1–12. https://doi.org/10.1145/3173574.3173650

[7] 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. https://doi.org/10.1145/1866029.1866080

[8] Jeffrey P. Bigham, Ryan S. Kaminsky, Richard E. Ladner, Oscar M. Danielsson, and Gordon L. Hempton. 2006. WebInSight: Making Web Images Accessible. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility* (Portland, Oregon, USA) *(Assets '06)*. Association for Computing Machinery, New York, NY, USA, 181–188. https://doi.org/10.1145/1168987.1169018

[9] Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P. Bigham. 2013. Visual Challenges in the Everyday Lives of Blind People. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2117–2126. https://doi.org/10.1145/2470654.2481291

[10] W3 Consortium. 2018. Web Content Accessibility Guidelines (WCAG) 2.1. https://www.w3.org/TR/WCAG21/

[11] Be My Eyes. 2023. Introducing Be My AI (formerly Virtual Volunteer) for People who are Blind or Have Low Vision, Powered by OpenAI's GPT-4. https://www.bemyeyes.com/blog/introducing-be-my-eyes-virtual-volunteer

[12] Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M. Kitani, and Jeffrey P. Bigham. 2019. "It's Almost like They're Trying to Hide It": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 549–559. https://doi.org/10.1145/3308558.3313605

[13] Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M. Kitani, and Jeffrey P. Bigham. 2020. Twitter A11y: A Browser Extension to Make Twitter Images Accessible. 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, 1–12. https://doi.org/10.1145/3313831.3376728

[14] Ricardo Gonzalez, Jazmin Collins, Shiri Azenkot, and Cynthia Bennett. 2024. Investigating Use Cases of AI-Powered Scene Description Applications for Blind and Low Vision People. arXiv:2403.15604 [cs.HC]

[15] Google. 2023. Get image descriptions on Chrome. https://support.google.com/chrome/answer/9311597?hl=en&co=GENIE.Platform%3DDesktop

[16] Darren Guinness, Edward Cutrell, and Meredith Ringel Morris. 2018. Caption Crawler: Enabling Reusable Alternative Text Descriptions Using Reverse Image Search. 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. https://doi.org/10.1145/3173574.3174092

[17] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P. Bigham. 2019. VizWiz-Priv: A Dataset for Recognizing the Presence and Purpose of Private Visual Information in Images Taken by Blind People. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[18] 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[19] Jaylin Herskovitz, Andi Xu, Rahaf Alharbi, and Anhong Guo. 2023. Hacking, Switching, Combining: Understanding and Supporting DIY Assistive Technology Design by Blind People. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 57, 17 pages. https://doi.org/10.1145/3544548.3581249

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

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

[22] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4565–4574.

[23] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[24] Klaus Krippendorff. 1970. Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement* 30, 1 (1970), 61–70. https://doi.org/10.1177/001316447003000105 arXiv:https://doi.org/10.1177/001316447003000105

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

[26] Tiffany Liu, Javier Hernandez, Mar Gonzalez-Franco, Antonella Maselli, Melanie Kneisel, Adam Glass, Jarnail Chudge, and Amos Miller. 2022. Characterizing and Predicting Engagement of Blind and Low-Vision People with an Audio-Based Navigation App. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 411, 7 pages. https://doi.org/10.1145/3491101.3519862

[27] Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding Blind People's Experiences with Computer-Generated Captions of Social Media Images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5988–5999. https://doi.org/10.1145/3025453.3025814

[28] Microsoft. 2023. Seeing AI. https://www.microsoft.com/en-us/ai/seeing-ai

[29] Vishnu Nair, Hanxiu 'Hazel' Zhu, and Brian A. Smith. 2023. ImageAssist: Tools for Enhancing Touchscreen-Based Image Exploration Systems for Blind and Low Vision Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 76, 17 pages. https://doi.org/10.1145/3544548.3581302

[30] OpenAI. 2023. GPT-4. https://openai.com/research/gpt-4

[31] Helen Petrie, Chandra Harrison, and Sundeep Dev. 2005. Describing images on the web: a survey of current practice and prospects for the future. *Proceedings of Human Computer Interaction International (HCII)* 71, 2 (2005).

[32] 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. https://doi.org/10.1145/3313831.3376404

[33] Abigale Stangl, Nitin Verma, Kenneth R. Fleischmann, Meredith Ringel Morris, and Danna Gurari. 2021. Going Beyond One-Size-Fits-All Image Descriptions to Satisfy the Information Wants of People Who Are Blind or Have Low Vision. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, Article 16, 15 pages. https://doi.org/10.1145/3441852.3471233

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

[35] Marynel Vázquez and Aaron Steinfeld. 2014. An Assisted Photography Framework to Help Visually Impaired Users Properly Aim a Camera. *ACM Trans. Comput.-Hum. Interact.* 21, 5, Article 25 (nov 2014), 29 pages. https://doi.org/10.1145/2651380

[36] Samuel White, Hanjie Ji, and Jeffrey P. Bigham. 2010. EasySnap: real-time audio feedback for blind photography. In *Adjunct 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, 409–410. https://doi.org/10.1145/1866218.1866244

[37] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. PMLR, 2048–2057.