ImageExplorer: Multi-Layered Touch Exploration to Encourage Skepticism Towards Imperfect AI-Generated Image Captions

Jaewook Lee
University of Illinois at Urbana-Champaign
Champaign, IL, USA
jaewook4@illinois.edu

Jaylin Herskovitz
University of Michigan
Ann Arbor, MI, USA
jayhersk@umich.edu

Yi-Hao Peng
Carnegie Mellon University
Pittsburgh, PA, USA
yihaop@cs.cmu.edu

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

ABSTRACT

Blind users rely on alternative text (alt-text) to understand an image; however, alt-text is often missing. AI-generated captions are a more scalable alternative, but they often miss crucial details or are completely incorrect, which users may still falsely trust. In this work, we sought to determine how additional information could help users better judge the correctness of AI-generated captions. We developed *ImageExplorer*, a touch-based multi-layered image exploration system that allows users to explore the spatial layout and information hierarchies of images, and compared it with popular text-based (Facebook) and touch-based (Seeing AI) image exploration systems in a study with 12 blind participants. We found that exploration was generally successful in encouraging skepticism towards imperfect captions. Moreover, many participants preferred ImageExplorer for its multi-layered and spatial information presentation, and Facebook for its summary and ease of use. Finally, we identify design improvements for effective and explainable image exploration systems for blind users.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; **Accessibility technologies**.

KEYWORDS

Automatic image captioning, alternative text, alt text, imperfect AI, touch exploration, screen reader, encourage skepticism, accessibility, Blind, visual impairment

ACM Reference Format:
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 *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3491102.3501966

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3501966

1 INTRODUCTION

Understanding images on the web can be challenging for blind or visually impaired (BVI) individuals. BVI users often depend on alternative text (also known as alt-text) [10] in order to understand the content of an image. However, in part due to the rapid increase in the quantity of user-uploaded content online, a growing number of images are missing alt-text, leaving a large fraction of images inaccessible [39]. While some platforms have provided users with the option to add alt-text as they upload a photo, these options are rarely used [53, 56]. For example, Gleason et al. found that only 0.1% of tweets with images contained alt-text [15]. To circumvent this problem, recent work has instead turned to auto-generating image captions [37, 51, 57] with the goal of providing high quality alt-text at scale. Automated systems have shown to greatly improve the coverage of alt-text [17, 20], but the quality and accuracy of these captions still remain questionable.

Prior work has noted that auto-generated captions are often error-prone, or missing key information about the image context, which has a negative effect on image understanding [44, 49]. When AI-generated captions are incorrect or misleading, without the means to verify correctness, BVI users place a high degree of trust in them, especially if they do not have access to additional information [35]. MacLeod et al. observed that BVI users often attempt to resolve discrepancies in captions by filling in details and developing their own reasoning that could explain the scenario [35]. To address this issue, they aimed to encourage skepticism in generated captions by altering the caption's framing.

Beyond being able to attribute errors to captions, enabling BVI users to identify specific errors in generated captions would provide a higher level of image understanding. One way to do this could be to provide richer ways to interact with an image beyond a single caption, so that users can investigate the auto-generated captions for themselves after gaining a better understanding of the image's content and layout. Prior work has suggested a variety of image exploration modalities, including image tags [12] and touch exploration [37, 38, 59]. Thus, in this work, we aim to address the following question:

> *What image exploration modality could best support BVI people in identifying errors in auto-generated image captions?*

To answer this question, we present an evaluation of three image exploration systems: Facebook's text-based 'Detailed Image Descriptions' feature [12], Seeing AI's touch-based image exploration feature [37], and an exploration system we developed as a design probe, *ImageExplorer*. ImageExplorer's design is inspired by Morris et al.'s finding that the following two approaches are helpful in improving BVI people's understanding of images: 1) providing alt-text through multiple "layers," where deeper layers contain additional detail, and 2) supporting touch-based interaction with images [38]. In this paper, we aim to combine these two approaches in a single system to provide as much information as possible, and compare it with state-of-the-art text- and touch-based image exploration systems to understand if doing so will give rise to skepticism towards auto-generated captions and further improve BVI users' abilities to identify errors in these descriptions.

Using these exploration systems, we conducted a comparison study with 12 blind participants. Participants explored a total of 9 images with varying caption qualities using the three systems and were asked to rate the accuracy of the auto-generated captions before and after explorations. After the participants have used all three systems, we asked them to rank the three systems based on ease of use, helpfulness, and overall preference, compare text- with touch-based systems, and compare single- with multi-layered systems. Specifically, we aimed to assess the following research questions:

RQ1: Do image exploration systems that provide additional information about an image help blind users better judge the correctness of AI-generated captions?
RQ2: Between text- and touch-based image exploration systems, which is more effective and why?
RQ3: Between the touch-based image exploration systems, is a single- or multi-layered approach more effective and why?
RQ4: What are users' perceptions and preferences towards image exploration systems?

We found that participants were unsure about the accuracy of auto-generated captions prior to exploring the images. After explorations, participants, on average, gave significantly lower accuracy ratings than their initial scores, which suggests elevated skepticism. When analyzing this further, there was a significant difference in change in scores for image caption qualities B (partially inaccurate) and C (inaccurate), but not A (mostly accurate), which indicates that participants were able to determine that an inaccurate caption is, in fact, inaccurate. Additionally, there was no significant difference in change in scores between the three systems overall and for each caption quality level. However, when more specifically comparing text- and touch-based systems, participants changed and decreased their score significantly more when using touch-based systems than the text-based one for images with inaccurate captions. Furthermore, ImageExplorer with multi-layered information led participants to have more correct explanations than Seeing AI and Facebook did. Overall, participants agreed that while a text-based system was easier to use, a touch-based system provided much more information about an image such as absolute and relative positions. When comparing a single-layered system with a multi-layered one, participants preferred the latter because it generates a hierarchy, which allows users to understand which main object that sub-objects belong to, and gives users the autonomy to choose whether to view the additional information or not. Finally, when asked which system they prefer overall, participants were split evenly between Facebook and ImageExplorer.

These results indicate that both text- and touch-based explorations encouraged skepticism towards imperfect AI-generated image captions. However, there were differences in user preferences, mainly, Facebook was the easiest to use, while ImageExplorer was the most helpful in understanding the content of the images. In summary, we contribute a thorough evaluation of image exploration modalities in allowing BVI users to judge caption accuracy. Our study revealed design improvements for effective and explainable image exploration systems in the future. Overall, our work demonstrates the potential of image exploration systems to allow BVI users to independently verify captions.

2 RELATED WORK

Image captioning is a widely researched area in accessibility. Specifically, our work builds on prior work in *(i)* image accessibility issues, *(ii)* automated image captioning systems, *(iii)* alternative image exploration systems, and *(iv)* explaining and understanding automated systems.

2.1 Image Descriptions

Image descriptions are the primary method for screen reader users to access image content online or in other software. Ideally, image descriptions are created by website authors following the Web Content Accessibility Guidelines [10]. However, a long stream of research has consistently found that this is not the case, with a recent estimate of 20-35% of images on top websites lacking image descriptions [5, 20, 40].

With the rise of social media and user-generated content, caption coverage has become significantly worse [53, 56]. In 2015, Morris et al. found that approximately 28.4% of English tweets contained some multimedia, and that in over 70% of these, the embedded images were considered important to understanding the meaning of the tweet [39]. Although Twitter now allows users to add alt text to their images, Gleason et al. found that only 0.1% of tweets with images contained alt text [15].

2.2 Automated Image Captioning

Given the lack of consistent captioning by content authors, a variety of automated approaches have been used to generate captions. These approaches generally fall into two categories, either using hybrid approaches such as crowdsourcing or web crawling to reuse captions, or advancing machine learning techniques to fully generate an image caption.

Hybrid methods for generating image captions have been used in prior work. A variety of crowdsourcing systems have been developed, and are generally successful in generating captions [4, 44]. For example, WebInSight provided a mechanism for users to request images on a web page to be sent to a labelling service for captioning [5]. Unfortunately, these systems are costly in price and latency. Caption Crawler used a different approach, they instead perform a reverse search for the image to scrape alt text from elsewhere on the web [20]. Sammani et al. build off of this approach by fetching existing captions and directly editing them with a language model [45]. While this works well for some online content, many images on social media are user generated and thus do not exist elsewhere.

Machine learning methods typically attempt to combine vision and language models in order to fully generate an image caption [13, 41, 51]. While many of these models were not designed to provide image descriptions for screen reader users, some systems have aimed to do this. For instance, Facebook's Automatic Alt-Text system originally aimed to generate image tags that describe the prominent objects in an image [57]. More recently, Facebook has updated this system to provide a full natural-language image description, along with providing tags grouped by position, prominence, and category [12]. Seeing AI similarly provides full image descriptions, and allows users to quickly obtain captions for photos locally on their mobile device [37]. Twitter A11y combines automated methods (optical character recognition and scene description) with hybrid methods (web crawling, link following, and crowdsourcing) to greatly increase the coverage of captions on Twitter [17].

However, prior work has noted that these generated captions are often error-prone, which has a negative effect on image understanding [44]. While Twitter A11y increases coverage greatly, they found that <60% of captions were high quality [17]. Even when technically correct, generated captions are often missing key information that users need to know to fully understand the context of why the photo was used or to decide if it is a good photo to post [49, 58]. MacLeod et al. found that blind users place a high degree of trust in automatically generated captions, and resolve dissonance by describing scenarios that would fit the caption. Specifically, they found that captions that emphasized the probability of error (i.e., "There's a small chance I'm wrong, but I think that's a cat sitting on a couch") encouraged more skepticism and caused users to attribute errors to captions more than positively-phrased captions [35]. In this case, encouraging skepticism in generated captions is potentially beneficial as it could allow users to better identify incorrect captions and thus have a better understanding of images overall. In this paper, we hope to further understand how to encourage skepticism in captions, and how blind users could independently identify errors in captions.

2.3 Alternative Image Exploration Systems

As an alternative to textual image descriptions, prior work has also explored using touch or multimedia systems to convey image content. Physical tactile image representations created by embossing, 3D printing, or tactile displays have been used to convey graphs, maps, and models [18, 19, 22, 27, 46, 48, 50]. Audio is commonly used alongside these representations or alongside typical captions to guide exploration [16, 46]. Prior work has tried to use common touch screens for a similar effect. Systems such as Seeing AI, RegionSpeak, and TouchCursor place bounding boxes over key objects in an image, then read out object descriptions as users move their finger into one of the bounding boxes [23, 37, 59]. These systems also use audio and haptic feedback to notify users when their finger enters or leaves a bounding box. Morris et al. evaluated a similar touch-based exploration system compared with two other alternative interactions for image understanding: providing layered captions or the ability to explore a single object further, and playing sound effects along with an image [38]. They found that touch exploration was promising in that it allowed users to visualize the spatial layout of an image, while providing layered captions was helpful in that users could choose what level of detail they wanted to know. Rastogi et al. similarly used a hierarchical approach to allow users to zoom into graphics on tactile displays [42].

In this work, we are interested in how touch-based exploration systems may be used to help users independently verify text captions. Given that touch-based exploration systems show potential for increasing image understanding, we investigate how users may resolve discrepancies between their visualization of an image from touch and the text caption. Additionally, building off of Morris et al.'s evaluation [38], we combine two promising approaches (touch exploration and layered captions) into one system to compare with existing approaches.

2.4 Understanding Automated Systems

With the rise of automated systems in all domains, research around transparent and explainable artificial intelligence has also grown. Doshi-Velez and Kim define interpretability as 'the ability to explain in understandable terms to a human.' Interpretability is used to confirm a variety of important factors in automated systems' decision making, including fairness, reliability, and trust [11]. While fairness and non-discrimination are aspects that are particularly relevant to the accessibility community [2, 14, 21], in this work we focus on how users perceive the reliability of an automatically generated image description.

Approaches to interpretability generally fall into two categories: creating inherently interpretable glass-box models, or providing post-hoc explanations for black-box models [29]. In image captioning, interpretability is typically achieved with post-hoc explanations that highlight regions of the image associated with a term or phrase in order to provide a visual reason for the chosen term [6, 24, 34, 43, 54]. Given these inherently visual explanations, alternative methods for increasing blind users' understanding of image captioning errors are needed.

As mentioned, one potential method for increasing skepticism in captions is to verbally frame them as uncertain [35], which is a form of model transparency [3]. While this increases awareness of potential errors, it does not necessarily help correct them. In this work, given that we would like to explain existing image captioning, we instead look at the potential of related models—character recognition, object recognition, and object segmentation—to serve as explanations. We thus compare text- and touch-based image exploration systems to understand their potential as proxies for post-hoc model explanations.

3 METHODS

In order to understand how different image exploration systems could support BVI users in judging caption accuracy and identifying errors in captions, we ran a user study with 12 blind iPhone users. During the study, we asked participants to use three systems—Facebook's text-based 'Detailed Image Descriptions' feature [12], Seeing AI's touch-based image exploration feature [37], and our touch- and layer- based ImageExplorer system—to obtain more information about an image and judge the accuracy of natural language captions generated by Microsoft Cognitive Services [9].

3.1 Design Probe: ImageExplorer

ImageExplorer is an image exploration system that uses touch and multiple information layers to allow users to explore the content of the image, their spatial relationships, and their hierarchy [31].

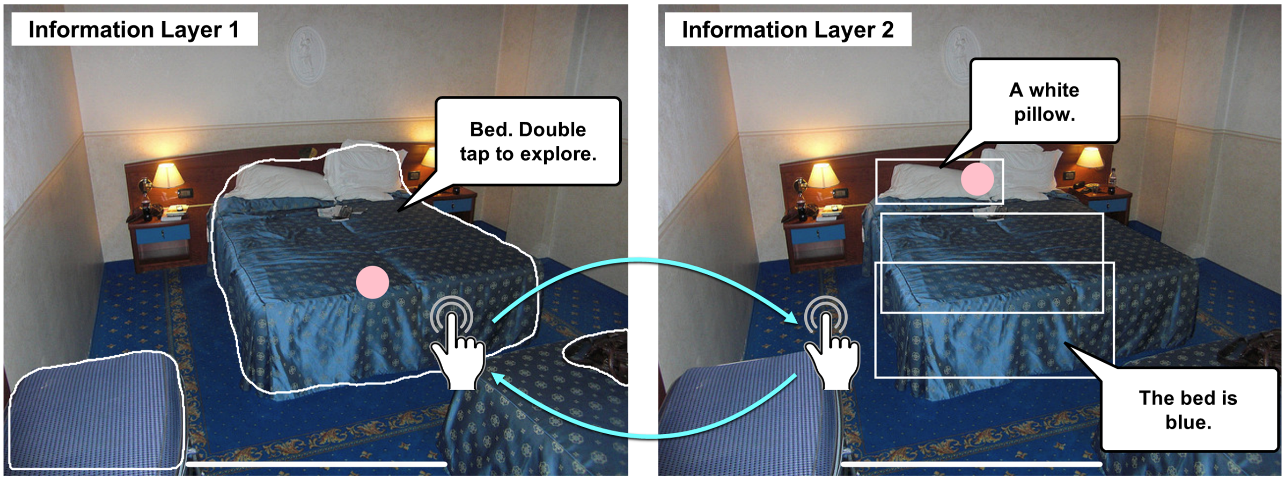


Figure 1: *ImageExplorer* user interface. The first information layer shows primary objects in the image outlined with polygonal boundaries. After double tapping on an object, users enter the second information layer, which shows rectangular bounding boxes around various detailed sub-objects. After exploring an object in detail, users can double tap anywhere to exit the second layer.

ImageExplorer was designed to allow users to identify common errors in auto-generated captions, including missing information, incorrect object labels, and incorrect layout descriptions (examples are shown in Figure 3). It is intended to supplement auto-generated natural language captions, which often have errors; we instead focus on providing BVI users with a variety of raw information from off-the-shelf models so that they can judge captions independently. It first collects information from an image through a handful of deep learning algorithms, aiming to reduce the probability of missing key information. It then separates this large quantity of information into two presentation layers using a set of criteria, allowing users to review details about an object and identify mislabeling. Finally, it provides a touch interface supported by audio feedback for accessing object information, allowing users to explore the spatial relationships between objects. We implemented ImageExplorer as an iPhone application due to pervasive use by the target population.

3.1.1 Element Detection and Scene Hierarchy. ImageExplorer leverages a variety of existing deep learning models to detect image content and create a scene hierarchy that can later be explored by users. It focuses on extracting common image elements: the location, boundaries, and descriptive labels of people and objects in an image, and transcriptions of printed text. Specifically, we used Mask R-CNN model [25] with ResNet-101 [26] and Feature Pyramid Network (FPN) [32] as backbone pre-trained on the MS-COCO dataset [33] to generate element masks and labels as the first layer of presented information. Compared with other object recognition models, Mask R-CNN is unique in that it generates segmentation masks, which are polygon-shaped borders that best fit elements of interest. Because ImageExplorer uses object boundaries to determine the scene hierarchy, tighter object borders resulted in a more accurate representation than traditional bounding boxes. Additionally, polygonal boundaries could better represent spatial aspects of an image such as size and shape of elements and overlap between elements.

To extract more regional and fine-grained information and descriptions needed to construct the second layer, we further utilized the existing Google Cloud Vision Model [8] to perform object, face and text detection and labeling, and the DenseCap model [28] with VGG-16 [47] as architecture pre-trained on the Visual Genome dataset [30] to produce more localized descriptions for specific image regions (e.g., "front wheel" and "back wheel" of a vehicle). Each regional element is displayed using a traditional bounding box if it was recognized with a confidence level of 75% or above. This threshold was chosen empirically based on our observations of performance, such that it removed many misleading labels while still maintaining a sufficient amount of labels overall. For instance, DenseCap described the street in image 1A in Figure 3 as "the tennis court is white" (70% confidence), which was removed by this threshold. On the other hand, the label, "the front wheel of the motorcycle" (85% confidence) was kept. This threshold is not flawless as it also removed a few correct labels such as "building is brick" (52% confidence) and kept incorrect labels such as "the shirt is white" (for the bee in image 2C, 87% confidence), but it neither removed too many labels nor kept many inaccurate labels.

To create a second layer in the hierarchical structure using the regional elements, we set up the following criteria: *(i)* the area of the regional element is smaller than that of the first-layer element, and *(ii)* at least 75% of the regional element overlap with a specific first-layer element. Both of these criteria were chosen because a bounding box may have areas outside the tighter polygon. The chosen second-layer elements were then paired with the first-layer element that they overlapped with the most.

Finally, any areas in both layers smaller than 34 pixels by 34 pixels were omitted to remove elements that are too small to touch. We chose this constraint based on two popular human interface design guidelines for mobile devices: Apple recommends a minimum constraint of 44 pixels by 44 pixels [1], and Microsoft suggests a minimum constraint of 34 pixels by 34 pixels [36]. To include as much information as possible, we opted to use the smaller constraint of the two. Note that these guidelines were not developed for people with visual impairments, which deserves future work.

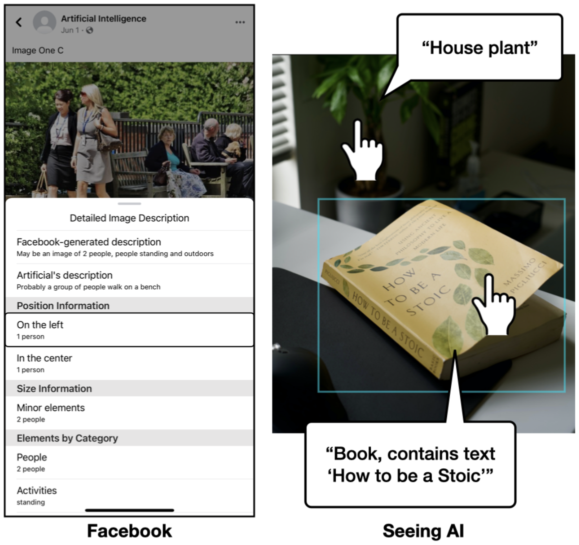
3.1.2 ImageExplorer User Interface. ImageExplorer provides a touch interface for exploring the content and hierarchy of an image that we extracted. An overview is shown in Figure 1. When the user opens an image, ImageExplorer first vocalizes the number of elements available, and displays the first-layer elements as polygonal boundaries overlaid onto the image for users to explore. As users move their finger across an image, they receive audio feedback: when not touching any element, a background tone plays; when touching an element, its name is read verbally (e.g., “bed,” “chair,” “handbag”). If an element contains sub-elements, users are then verbally prompted to double tap for more information (e.g., “bed, double tap to explore”). If the user chooses to double tap on a first layer element, the system will display its corresponding second layer elements, which the users can again explore using touch (e.g., “a white pillow” and “the bed is blue”). Users can exit the second layer at any time by double tapping anywhere on the screen. When the user returns to the first layer, the system will say the number of elements yet to be explored, providing awareness of exploration progress (e.g., “going back to the whole image, two objects remaining”). If a first layer element does not have any second layer elements, the system will decrement the number of elements left to explore as soon as the user touches that element and provide an updated number to the users.

3.2 Participants

We recruited 12 BVI participants from an emailing list. Our study was approved by the university’s Institutional Review Board, and participants consented to participation, screen and audio recordings through both email and verbal responses. At the start of the study, we asked participants for their demographic information (Table 1). Participants were between 27 and 69 years of age (μ = 48.17 years, σ = 14.69), with six being female and six being male. Seven were totally blind (four of them were born blind), while the rest were legally blind with some light perception. When asked how familiar they were with Facebook, they rated an average of 5.58 out of 7 (σ = 1.62, where 5 is somewhat familiar, and 6 is familiar). When asked how familiar they were with Seeing AI, they rated an average of 5.42 out of 7 (σ = 2.02). All of our participants reported that they use VoiceOver as their mobile screen reader.

ID	Gender	Age	Vision Level
P1	Female	66	Some light perception, since age 10
P2	Female	46	Fully blind, since age 20
P3	Male	45	Light perception, since childhood
P4	Female	67	Fully blind, since birth
P5	Male	69	Fully blind, since birth
P6	Male	27	Some light perception
P7	Female	38	Fully blind, since birth
P8	Male	34	Fully blind, since childhood
P9	Female	58	Fully blind, since birth
P10	Male	30	Some color perception in peripheral vision
P11	Male	43	Fully blind, since age 31
P12	Female	55	Some light perception, since birth

Table 1: Participant demographics for our user study.



Artificial Intelligence
Jun 1 ·
Image One C

Detailed Image Description

Facebook-generated description
May be an image of 2 people, people standing and outdoors

Artificial's description
Probably a group of people walk on a bench

Position Information

On the left
1 person

In the center
1 person

Size Information

Minor elements
2 people

Elements by Category

People
2 people

Activities
standing

Figure 2: Facebook and Seeing AI’s exploration interfaces: Left: Facebook’s detailed image description interface. A list of text grouped into categories. Users can swipe to read through the textual information. Right: Seeing AI’s touch exploration interface. When entering an element, it reads the name of that element.

3.3 Apparatus

We asked participants to download the Facebook, Seeing AI and ImageExplorer applications onto their phones prior to the study session. The study was conducted remotely using the Zoom mobile app due to the COVID-19 pandemic. Participants shared the screens and audio of their devices as they completed the study. Thus, study coordinators could both verbally provide instructions to participants and observe how they used each application.

For Facebook, we set up a new account with posts containing the images used for our study and asked participants to log in using our credentials prior to the study. To explore an image using Facebook’s Detailed Image Descriptions feature, participants first navigate to a post containing that image, and Facebook will read its alt-text. To control for the study, we replaced Facebook’s auto-generated alt-text with the ones generated by Microsoft Cognitive Services [9]. Participants can swipe up or down on the post until they hear “Generate Detailed Image Descriptions” and then double tap to activate it. Facebook presents additional information about an image as a list of text, which is grouped into multiple categories such as Position Information, Size Information, and Elements by Category (an example is shown in Figure 2). Participants can access this information using screen reader gestures.

For Seeing AI, we sent participants the necessary images in an email thread so that they could download them prior to the study. To explore an image using Seeing AI, participants first choose Browse

Photos and select the image to explore. Seeing AI will read the auto-generated caption pertaining to that image. Participants can then activate the Explore button to enter the touch exploration interface, where objects are presented as single-layered bounding boxes that participants can move their finger along the interface to hear real-time feedback of what is underneath their finger (an example is shown in Figure 2).

The ImageExplorer iOS app was distributed using TestFlight, which we asked participants to install prior to the study. As described in Section 3.1, participants can explore an image using touch similar to that in Seeing AI. In addition, participants can double tap on a first layer element to access its corresponding second layer elements (an example is shown in Figure 1).

3.4 Image Selection

To pick the images used in our study, we first randomly selected and generated captions for a total of 30 images from MS-COCO [33] and Unsplash [52] image data sets. We then classified the automated caption accuracy following a similar process to Gleason et al. [15, 17], who used four quality levels for human-written alt text: irrelevant, somewhat relevant, good, and great. However, no auto-generated captions were 'great' and even the best ones are still missing secondary or minor information and could be improved by providing additional details [49]. Therefore, we modified these ratings to instead use three quality levels: mostly accurate (A), partially inaccurate (B), and inaccurate (C). Caption quality level A was assigned when an image had a mostly accurate caption, but is missing minor information and could be slightly improved. Caption quality level B was assigned when an image had a somewhat inaccurate caption, for example, if an object was missing a label, had an incorrect count, or if a single object was labeled incorrectly. Captions of this level do not necessarily detract completely from someone's understanding of the content. Caption quality level C was assigned when a caption was very inaccurate, where the labels or situation in the image were described completely incorrectly. These errors are significant because they detract completely from someone's understanding of the image's content and meaning.

Finally, to limit the study duration, we chose three images of each quality, for a total of nine images (shown in Figure 3). To do so, we selected images with varying subjects (e.g, people, animals, object), scenes (e.g., indoor, outdoor), and scales (e.g., close or wide shot). The nine images were then grouped into three sets such that each set contains one mostly accurate (A), one partially inaccurate (B), and one inaccurate caption (C). For example, image set 3 contains: 3A, which represents an image with a correct caption —"Probably a book on a table," 3B, which represents an image with a partially correct caption —"A cat sitting on a chair," although the cat is on a table, and 3C, which represents an image with an incorrect caption —"Probably calendar," although the image is about a bag of Kraft mozzarella cheese.

3.5 Procedure

Participants used each of the three systems to explore a different set of three images with three quality levels as described above. The order in which participants used each system was randomized and counter-balanced, as was the order of the given image set. Over the course of the study, participants thus explored all nine images. Participants explored three of the nine images for each system. For example, P1 first used ImageExplorer to explore images 2C, 2A, then 2B; then used Facebook for images 3A, 3B, and 3C; finally, they used Seeing AI for images 1B, 1C, and 1A.

For each system, participants first explored a tutorial image with a correct caption to familiarize themselves with the system. Then for each study image, participants were first provided with its auto-generated caption from Microsoft Cognitive Services [9], and we asked them to rate their agreements with the statements 'This caption is accurate,' and 'I am confident in my scores' on a 7-point Likert scale (where 1 is 'strongly disagree' and 7 is 'strongly agree'). Then, participants explored the image using one of the three systems and rated it again with the same set of questions. Once participants finished exploring the three images using each system, we asked them to rate their agreement with the statements 'This system was easy to use' and 'This system was helpful.' For each statement, we also asked them to explain their reasoning for their rating. Finally, after participants used all three systems, we asked them to: rank the three systems in terms of *(i)* ease of use, *(ii)* helpfulness in understanding the images, *(iii)* preference of use, as well as *(iv)* compare Facebook with the two touch-based systems (i.e. Seeing AI and ImageExplorer) and *(v)* compare Seeing AI with ImageExplorer. The study took about two hours, and participants were each compensated for $50. The screen and audio recordings were collected and transcribed for further analysis.

3.6 Analysis

We adopted a mixed-methods approach and performed both quantitative and qualitative analysis on our data. We first define the factors and measures in our quantitative analysis. The main measurement we analyzed was the accuracy ratings of image captions (1 = "very inaccurate", 7="very accurate") and their changes. To understand whether participants can identify image caption quality before exploration, we took caption quality as the factor (A, B, C), and compared the measurement of raw caption accuracy score using one-way ANOVA with follow-up Tukey's HSD post-hoc tests. Regarding the exploration time, we compared the measurement of exploration duration across three systems (as factors) using the same statistical analysis.

For the post-exploration analysis, the next step was to understand whether participants change their initial accuracy ratings and have better judgement on caption correctness after any form of exploration (RQ1). We took completion of exploration as the factor (before vs. after), and compared the measurement of raw caption accuracy scores with respect to *(i)* all images and *(ii)* each image set with different caption quality using Student's t-test (two-tail). Following that, we further explored if different exploration methods changed the caption accuracy scores differently for *(i)* all images, and *(ii)* each image set with different caption quality (RQ3). We took image exploration approaches as the factors (Facebook vs. Seeing AI vs. ImageExplorer) and compared the changes in accuracy scores using one-way ANOVA with follow-up Tukey's HSD post-hoc tests. Besides differences between three systems, we also want to understand the specific differences between touch- and text-based explorations (RQ2). We took interaction modalities (touch vs. text)

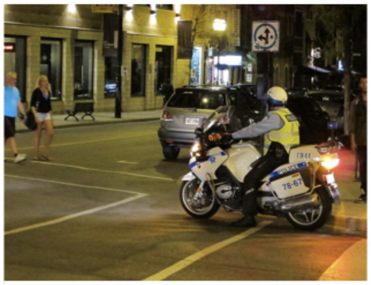
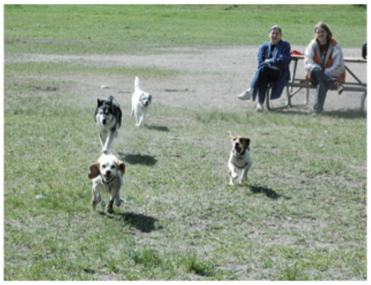
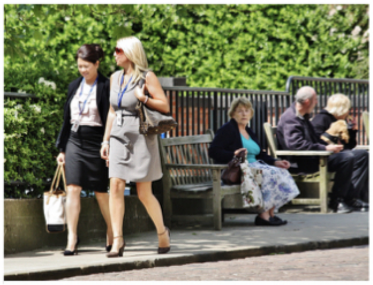
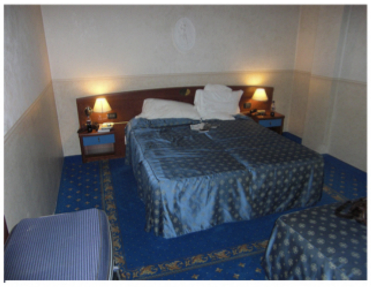
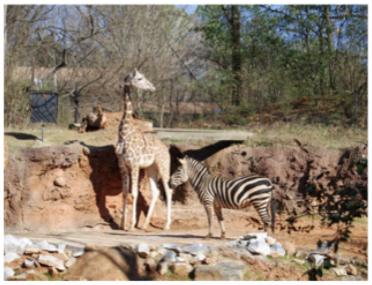
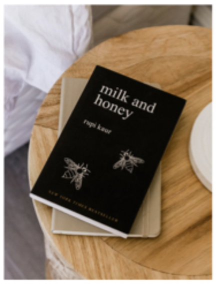
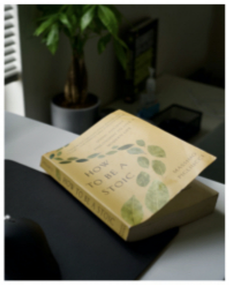
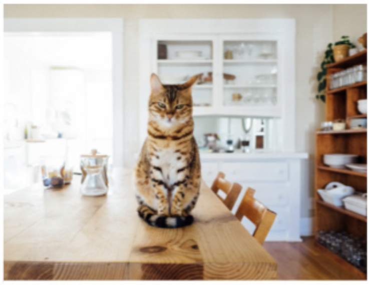
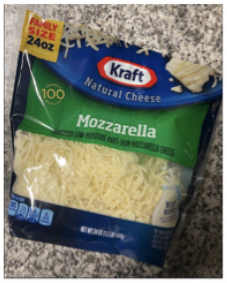
	Caption Quality A: **Mostly Accurate**	**Caption Quality B:** **Partially Inaccurate**	**Caption Quality C:** **Inaccurate**
Image Set 1	 **1A:** “A police officer on a motorcycle.” **Error:** Missing minor background information (people and cars)	 **1B:** “A group of dog running on grass.” **Error:** Missing major background information (people and bench); incorrect grammar	 **1C:** “Probably a group of people walk on a bench.” **Error:** Incorrect activity labels (walking and sitting
Image Set 2	 **2A:** “Probably a bed with blue sheets.” **Error:** Missing minor background information (pillows, lamps, chair)	 **2B:** “A giraffe and zebras in a zoo exhibit.” **Error:** Incorrect counts (singular giraffe and zebra)	 **2C:** “A black box on a table.” **Error:** Incorrect label (book)
Image Set 3	 **3A:** “Probably a book on a table.” **Error:** Missing minor background information (mouse, plant, desk)	 **3B:** “A cat sitting on a chair.” **Error:** Incorrect object relationships (cat on table)	 **3C:** “Probably calendar.” **Error:** Incorrect label (bag of cheese)

Figure 3: Images used during the user study sessions, their AI-generated captions, and the errors in those captions.

as factors and compare the changes in accuracy scores for *(i)* all images, and *(ii)* each image set using Student’s t-test (one-tail). We used one-tail analysis because we hypothesized that touch is more effective than text-based explorations.

Besides caption accuracy ratings, we also review the differences on the level of participants’ confidence on their judgements of the ratings (1 = “very unconfident”, 7=“very confident”). For pre-exploration stage, we took caption quality as factor (A, B, C) and compared the measurement of confidence level using one-way ANOVA with follow-up Tukey’s HSD post-hoc test. In addition to investigating if caption quality affect their confidence level before exploration, we also took completion of exploration as factor (before vs. after) and compare the measurement of confidence level using Student’s t-test (two-tail), in order to understand if any form of exploration affect the confidence of their determination on caption accuracy. Note that the alpha level of all our conducted tests was 0.05. We took 7-point Likert scale as approximating equal intervals and thus analyzed them using ANOVAs or t-tests. Overall, our results were consistent when validated with non-parametric tests.

For our qualitative analysis, two members of the research team analyzed the study sessions using thematic analysis as described by Braun and Clarke [7]. We first created written descriptions of

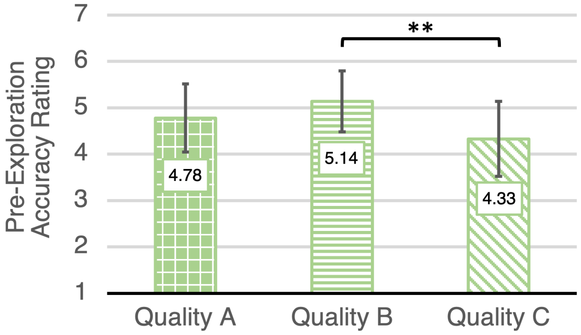


Figure 4: Pre-exploration accuracy ratings for each image quality level. Participants gave significantly higher ratings for images of quality B as compared to images of quality C.

participants' app usage behavior from the study video recordings, e.g., how often they re-read information or how they used touch to explore an image. These behavioral descriptions, along with study transcripts, were treated as data items to identify trends in participant feedback. We first individually read and familiarized ourselves with the data. We performed an open coding of the data independently, then adjusted the codes as a group until sufficient agreement was reached. We focused on identifying themes relating to participants' exploration strategies, accuracy rating reasoning and image interpretations, and overall app preference reasoning.

4 RESULTS

4.1 Pre-Exploration Caption Accuracy Ratings

We first analyzed the initial perceived accuracy scores (1-7, 7 means very accurate) that participants gave to the generated captions before explorations (Figure 4). For images with captions of quality levels A (Mostly Accurate), B (Partially Inaccurate), or C (Inaccurate), participants gave the following average ratings: $\mu_A = 4.78$ with $\sigma_A = 1.47$, $\mu_B = 5.42$ with $\sigma_B = 1.11$, and $\mu_C = 4.06$ with $\sigma_C = 1.58$. Our statistical results showed that there were significant differences between the accuracy ratings among three quality levels ($F(107) = 8.235; p = 0.0005 < 0.05$). Specifically, participants gave significantly higher ratings on caption quality B than C ($p = 0.001 < 0.05$), but no difference was found between A vs. B ($p = 0.14$) and A vs. C ($p = 0.08$). For the confidence level on the ratings they provided, our results showed that participants held similar confidence on their judgements regardless of caption quality ($F(105) = 0.064; p = 0.94$), with relatively high level of confidence on image set A ($\mu_A = 6.09; \sigma_A = 1.00$), B ($\mu_B = 6.06; \sigma_B = 0.97$) and C ($\mu_C = 6.00; \sigma_C = 1.04$). Overall, the average pre-exploration accuracy ratings were not high (from 4.06 to 5.42) and did not fully reflect the ground-truth quality groupings, indicating that while participants attempted to rate the captions' accuracy as well as possible, ultimately, they were unsuccessful in determining which captions were accurate and which were not. This suggests that information beyond just the caption is necessary to empower blind users to judge the correctness of AI-generated captions.

Participants' strategies for rating the generated captions generally fell into two categories: *(i)* analyzing the grammar of the captions, and *(ii)* judging captions based on prior knowledge.

4.1.1 Caption Grammar. Participants often used grammatical components of the caption to judge its accuracy. For example, the use of the word "probably" in captions caused participants to express skepticism towards the accuracy of those captions: *"It already says 'probably.' Given how the system is not sure, how can I be?"* (P11). Microsoft Cognitive Services [9] includes the word "probably" in captions where it has lower confidence, which was common in our image set (in four (1C, 2A, 3A, and 3C) of the nine images). This is consistent with prior findings by MacLeod et al. [35]. Additionally, grammatical errors in captions caused participants to view them as less accurate. For example, the caption for image 1B is "A group of dog running on grass." This confused many, including P8, who said *"Maybe it is nitpicking grammar thing, but it should be 'a group of dogs.' Now I am not sure if there are many dogs or just one dog."*

4.1.2 Prior Knowledge. Participants also relied on their prior knowledge of the world, AI, Facebook, and Seeing AI when initially rating the captions. Captions that seemed realistic generally were perceived as more accurate. For instance, image 2C is captioned "a cat sitting on a chair," which P1 described as likely accurate because: *"Cats do sit on a chair. When I first heard the caption, I thought of a cat sitting on top of an easy chair."* Likewise, captions that did not sound reasonable were perceived as less accurate. For example, image 1C is captioned "probably a group of people walk on a bench," which P3 described as inaccurate because: *"you can't really walk on a bench. I can sit on a bench, but not walk on a bench."*

Participants also occasionally used their prior experiences with AI-generated captions in general to reason about accuracy. Participants rated captions as correct if they believed it contained a distinct object that would be difficult to mistake for something else. For example, P3 noted: *"Giraffes and zebras are so distinct, so it couldn't be confused with something else, like a dog or a cat."* Additionally, participants who had experience using either Facebook or Seeing AI were slightly biased, and occasionally rated captions based on the quality of their prior experiences. For example, P4, who uses Facebook every few days, said *"Most of the time Facebook gives good captions."*

4.2 Exploration Time And Strategies

Our results demonstrated that there was a significant difference for users' exploration time among three different systems ($F(107) = 33.07; p < 0.0001$), shown in Figure 5. Specifically, participants spent significantly more time on ImageExplorer than the two other systems ($\mu = 168.33s$, $\sigma = 89.83; p = 0.001 < 0.05$). Participants also spent significantly more time on Seeing AI ($\mu = 87.83s$, $\sigma = 57.33s$) than Facebook ($\mu = 48.69s$, $\sigma = 28.26$) ($p = 0.027 < 0.05$). By analyzing participants' strategies when exploring the images using each of the three systems, we found that when using Facebook, all twelve participants first read through the textual information from top to bottom. Once they reached the end of the list of text, four of them listened to all of the information again, this time from bottom to top, while the rest did not explore further. On the other hand, when using the two touch-based systems (i.e., Seeing AI and ImageExplorer), most participants moved randomly without a strategy, with only two participants occasionally moving more strategically in a circular or zig-zag motion.

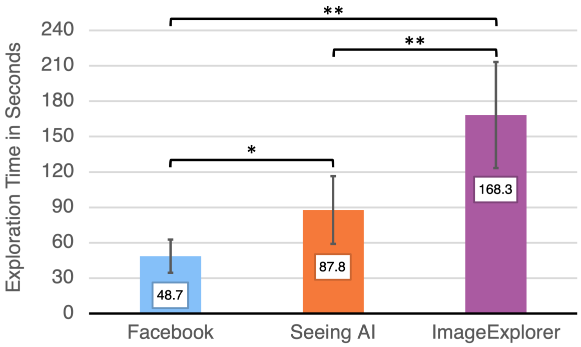


Figure 5: Average image exploration time when using Facebook, Seeing AI, and ImageExplorer.

While a lack of strategy could partially explain the difference in exploration time, the touch-based systems Seeing AI and ImageExplorer both provide additional information (e.g., object location and hierarchies) that take more time to explore. Additionally, some participants including P3 mentioned that with the touch-based systems, they tried to build up a picture of the image in their head: *"... with touch, you can get a little bit of information about what's where, and make a more accurate mental picture."* Though this approach potentially helps with error identification, it could also cause the increase in exploration time, as it increases the mental load of exploration.

4.3 Determining Caption Correctness

4.3.1 Effects of Exploration on Caption Judgement. To answer RQ1, we first need to know whether image exploration can affect participants' judgement on image caption accuracy regardless of both the quality of the captions and the systems used. To achieve this, we compared all of the pre-exploration accuracy ratings ($\mu = 4.75$; $\sigma = 1.51$) with all of the post-exploration accuracy ratings ($\mu = 3.39$; $\sigma = 1.97$). The results showed that there was a significant change in accuracy ratings ($t(212) = 5.65$; $p < 0.0001$), which indicates that explorations of any kind made participants change their initial accuracy scores regardless of any caption quality (Figure 6). Regarding if exploration changed the confidence level on their accuracy judgement, our results showed that participants did not really change the confidence of their own judgement when comparing the ratings before ($\mu = 6.05$; $\sigma = 1.00$) and after ($\mu = 6.19$; $\sigma = 1.00$) exploration ($t(212) = 1.02$; $p = 0.31$).

We then analyzed whether exploration changed participants' accuracy ratings for each image caption quality, regardless of the system used. To accomplish this, we first separated both the pre- and post-exploration data into three chunks based only on image caption quality, each containing accuracy ratings for image quality level A ($\mu_{pre} = 4.78$; $\sigma_{pre} = 1.47$; $\mu_{post} = 4.20$; $\sigma_{post} = 1.72$), B ($\mu_{pre} = 5.14$; $\sigma_{pre} = 1.32$; $\mu_{post} = 3.31$; $\sigma_{post} = 1.98$), and C ($\mu_{pre} = 4.33$; $\sigma_{pre} = 1.62$; $\mu_{post} = 2.67$; $\sigma_{post} = 1.90$). We found that there was a significant difference in scores before and after explorations for caption qualities B ($t(70) = 4.37$; $p < 0.0001$) and C ($t(69) = 4.91$; $p < 0.0001$), but there was no significant difference in scores for A ($t(69) = 1.50$; $p = 0.07$), shown in Figure 6. This result indicates that image exploration systems did help blind users better judge the correctness of captions, thus significantly decreased their scores for images with lower quality captions (B and C) but did not significantly change their scores for images with higher quality captions (A) after exploration.

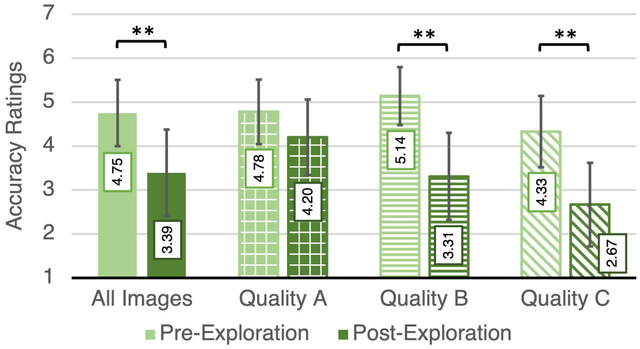


Figure 6: Pre- and post-exploration accuracy ratings for all images, and for images of each quality level.

4.3.2 Effects of Each System on Caption Judgement. We then examined for each system, whether explorations affect participants' accuracy ratings regardless of caption qualities. On average, participants changed their ratings the least when using Facebook's text-base explorations ($\mu = -1.11$; $\sigma = 1.31$). When comparing the two touch-based systems, participants, on average, changed their ratings less when using ImageExplorer ($\mu = -1.34$; $\sigma = 2.14$) than Seeing AI ($\mu = -1.60$; $\sigma = 1.57$). However, there was no significant difference in change of accuracy ratings between the three systems statistically ($F(105) = 0.71$; $p = 0.49$). Still, it is interesting to note that no matter which system participants used, they, on average, rated the accuracy of the captions to be lower after exploring.

We further analyzed for each caption quality, whether exploration with each of the three systems changed participants' accuracy ratings differently. We report the average change in accuracy ratings with standard deviation in parentheses as follow:

- **Image Caption Quality A**: Facebook: -0.67 (1.07); Seeing AI: -1.00 (1.81); ImageExplorer: 0.00 (1.61).
- **Image Caption Quality B**: Facebook: -1.42 (1.51); Seeing AI: -1.58 (1.08); ImageExplorer: -1.75 (2.34).
- **Image Caption Quality C**: Facebook: -1.33 (1.30); Seeing AI: -2.36 (1.63); ImageExplorer: -2.17 (1.99).

Here, we again computed and compared the changes in scores, rather than the raw scores. Results indicate that there was no significant difference in change of accuracy ratings across the three different systems for the three image caption quality levels A ($F(34) = 1.27$; $p = 0.30$), B ($F(35) = 0.11$; $p = 0.89$), and C ($F(34) = 1.26$; $p = 0.30$). However, the number of accurate explanations as to why the caption is correct or incorrect differed across the three systems. We define a correct explanation as a reasoning that describes which part of the caption is correct or incorrect and the image content with high accuracy (e.g., P3 said *"Well, because it's a book and not a box. But a book can be in the shape of a box, so."* after exploring image 2C). The three systems elicited the following number of correct explanations from our participants (Figure 7):

- **Facebook**: A: 3/12 B: 1/12 C: 1/12
- **Seeing AI**: A: 3/12 B: 7/12 C: 5/11
- **ImageExplorer**: A: 6/11 B: 8/12 C: 6/12

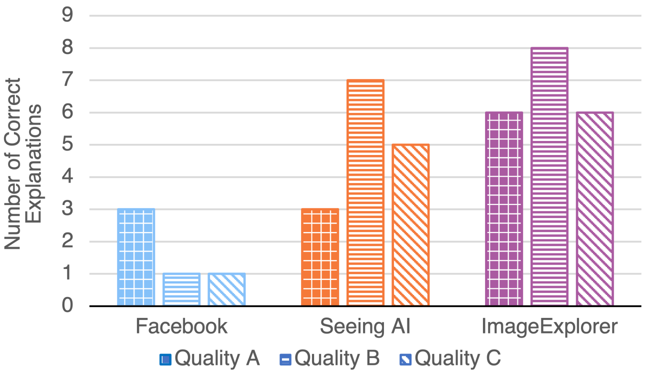


Figure 7: Number of correct explanations participants were able to give about image content, with each system and for each image quality level.

By reviewing the number of correct explanations, it is apparent that Seeing AI did better than Facebook, while ImageExplorer did better than both. Facebook elicited the least number of correct explanations due to a lack of detail. For instance, for image 3B, Facebook did not provide any information about the chairs or the table, which led participants to assume that the cat is sitting on a cabinet. Seeing AI provided spatial information, which P6 used to find out that the cat is sitting on top of a dining table and not a chair in image 3B. ImageExplorer not only provided spatial information, but also provided much more fine-grained information than Seeing AI, allowing P11 to describe image 2A as *"An image of a bed in the middle of a room with a white pillow and blue sheets with one chair to its left and one handbag on the floor to its right."*

4.4 Touch- vs. Text-Based Exploration

To answer RQ2 and better understand the differences between touch- and text-based image exploration systems, we further conducted both quantitative and qualitative analysis to directly compare the two types of systems. Quantitatively, our results show that for image caption quality C (inaccurate captions), the touch-based systems made participants decrease the accuracy ratings significantly more ($\mu = -2.26$, $\sigma = 1.55$) than the text-based system ($\mu = -1.33$, $\sigma = 1.25$) ($t(34) = 1.6$; $p = 0.044 < 0.05$), indicating that touch interactions helped raise participants' skepticism towards incorrect captions. However, for images with partially inaccurate captions (image caption quality B), the touch-based systems did not make participants change their accuracy scores ($\mu = -1.67$, $\sigma = 1.04$) significantly when compared with the text-based system ($\mu = -1.42$, $\sigma = 1.44$) ($t(34) = 0.37$; $p = 0.34$). Similar results could also be found for images with mostly accurate captions (image caption quality A) when comparing the change in scores of the touch-based systems ($\mu = -0.52$, $\sigma = 1.74$) with the text-based system ($\mu = -0.67$, $\sigma = 1.03$) ($t(33) = 0.26$; $p = 0.38$).

Participants also described the pros and cons of both text-based and touch-based systems. Participants who found the text-based system helpful reasoned that *(i)* it is quick and easy to use because it relies on the swipe gesture (5/12 participants), *(ii)* it takes less effort to locate the information (2/12), and *(iii)* there is a smaller chance of missing an available information (2/12). Those that disliked the text-based system said *(i)* text cannot provide spatial information such as absolute and relative locations, and size information (5/12), and *(ii)* text provides very little information overall (4/12). On the other hand, participants who liked the touch-based systems reasoned that *(i)* it provides spatial information such as absolute and relative locations, and size information (5/12), *(ii)* it provides a lot of information (3/12), and *(iii)* it promotes a sense of autonomy (3/12). Those that disliked the touch-based systems said *(i)* touch takes longer (2/12) and *(ii)* it is difficult to locate all of the elements in an image (2/12).

4.5 Single- vs. Multi-layered Exploration

For touch-based image exploration systems, we sought to understand the differences between presenting information in a single layer (e.g., Seeing AI) and multiple layers (e.g., ImageExplorer) (RQ3). As shown in Section 4.3, there was no quantitative difference between the change in scores of all systems. Even though participants change their ratings on caption quality similarly, as pointed out in Section 4.3.2, ImageExplorer did enable more correct explanations (20/35) than Seeing AI (15/35), implying that ImageExplorer might have a better chance to empower users to make better judgements on the actual caption quality. Furthermore, participants' provided feedback demonstrated notable differences between the two touch-based systems.

Specifically, half of the participants expressed that ImageExplorer with multi-layer hierarchy provided more information than Seeing AI with only single-layer descriptions. For instance, P6 found it helpful to know that the table is a wooden table in image 3B. On the other hand, P5 commented that sometimes ImageExplorer provides too much information. For example, it might not be necessary to know the exact number on the motorcycle in image 1A. While providing a lot of detail seem to help participants judge the caption accuracy, balancing between providing a lot of information and filtering out useless information still remains a challenge. Besides additional information, nine of the participants also found the hierarchical presentation coupled with double tap functionality useful since it not only helped them gain additional information in an organized way, but also gave users the autonomy to choose whether to view the additional information. For example, P12 noted: *"I like going into each of the objects. By having a description and then digging down to get more information, the information isn't crowded. For example I want to know more about the book, but not the plant."* While hierarchies seem effective, the interactions may need to be carefully design to reduce users' mental effort. For instance, P4 particularly disliked double tapping to access the second layer, because *"it was just one extra step that you had to go through."* P6 meanwhile suggested that using touch and hold gesture could potentially reduce both physical and mental effort for the users.

Besides the differences between single- and multi-layered exploration, participants also mentioned some specific feature differences between ImageExplorer and Seeing AI including (i) explored element indication, and (ii) responsiveness. Three participants found the feedback for whether an element was explored helpful since it allowed them to focus on navigating the unexplored area. P9 noted: *"It was helpful to know the amount of objects, and if it was the same object. So I didn't have to wonder if it was the same bench"*, and

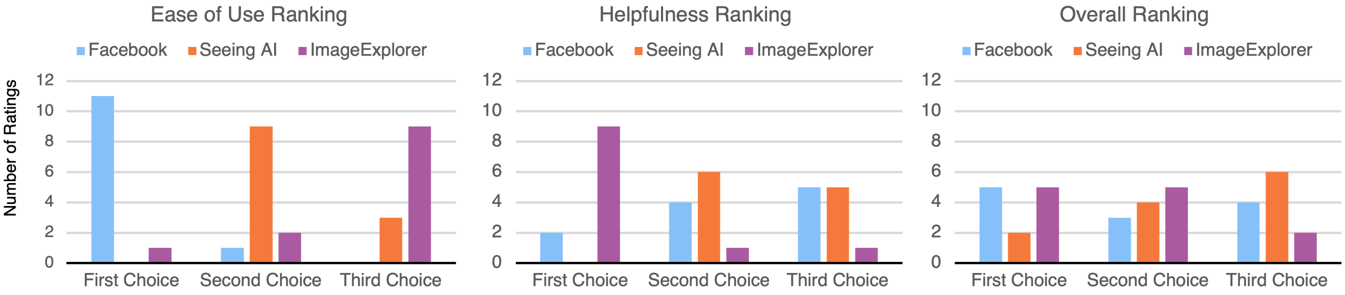


Figure 8: Post study rankings of Facebook, Seeing AI, and ImageExplorer for ease of use, helpfulness, and overall preference.

they expressed disappointment towards Seeing AI for not having this feature: *"...it doesn't tell you if you're back in an area you've already explored, you just have to know based on touch and spatial layout, which is difficult."* With regard to the responsiveness of the system, three participants felt Seeing AI was more responsive than ImageExplorer due to the differences in their audio feedback design. Specifically, Seeing AI plays a continuous melody when not touching any element and plays a loud "ding" when entering an element. On the other hand, ImageExplorer plays a monotone when not touching any element and does not play a loud "ding" when entering an element. This was done in an effort to provide the same information in a more subtle manner, but might lower the perceived responsiveness of ImageExplorer. As described by P8, *"The challenge with ImageExplorer is knowing whether or not it is working because of lack of feedback on empty space."*

4.6 System Preferences

As shown in Figure 8, 11 out of 12 participants found Facebook to be the easiest to use. This is because it is easier and faster to retrieve information from scrolling through a list of text than dragging a finger across the screen hoping to find an element. This consensus is best summarized by P3: *"I think just the quickness from Facebook is good. I was just able to scroll through text and that gave me all the context I needed for the photo."* Additionally, P5 supplemented this explanation by saying: *"It was frustrating to explore the screen hoping to land on something. It almost became a game to find the other elements in the image. I became more concentrated in finding all of the things in the image than understanding the image."* Findings from the qualitative data indicates that using a text-based system is more efficient than using a touch-based system, thus was the easiest to use.

On the other hand, 9 out of 11 participants (omitted one participant because they rated all three systems to be equal) found ImageExplorer to be the most helpful when judging the accuracy of the auto-generated captions. All of them agreed that ImageExplorer provides the most amount of information, including P10, who said *"ImageExplorer was the most detailed, so it was helpful. Facebook and Seeing AI are about the same, but Facebook is easier to use. Both seem to only capture the main elements in the image,"* and P7, who said *"Facebook and Seeing AI are not detailed enough."* Here, detailedness of the systems tend to correlate to their helpfulness. Additionally, 6 of those 9 participants appreciated ImageExplorer's hierarchy and double tap features because these features *(i)* showed which sub-elements belong to which main elements in the image, and *(ii)* broke down complex images into smaller, more manageable chunks. For instance, P9 found ImageExplorer to be most helpful because *"It allowed you to see what things were part of another thing, like that it was the bus's window and not some other window."* Additionally, P11 said *"The good thing about ImageExplorer was that you had a hierarchy to it. Objects that are complex can be divided into categories. It would not read the entire image, but focus on just one object. This is good especially for large images or complex images. The depth of layers is a research question. I don't suggest 7 hierarchies, but the 2 level hierarchy was good. I knew I was looking at just one object."* The collected qualitative data indicates that the participants found ImageExplorer to be the most helpful because it provided a lot of information in a structured and hierarchical manner.

Furthermore, we asked participants to provide a ranking for their overall preferences of the three systems. Out of the 12 participants, 5 chose Facebook, 5 chose ImageExplorer, and 2 chose Seeing AI (see Figure 8). Participants who preferred Facebook prioritized ease of use over detailedness, while those who preferred ImageExplorer thought the opposite. This suggests that blind users need both ease of use and detailedness when exploring an image. As a solution to this issue, 7 out of 12 participants suggested a system that merges features of Facebook and ImageExplorer, including P12, who specifically stated: *"If I can get my Facebook overview in ImageExplorer, that would be my favorite."* Finally, Seeing AI was ranked the lowest by 6 out of the 12 participants because it was not as easy to use as Facebook, but also not as detailed and helpful as ImageExplorer. These two characteristics seem to be the main factors that determined users' system preferences.

5 DISCUSSION AND FUTURE WORK

Our work demonstrates the potential for providing additional and structured information to help BVI users have better image understanding. We include a reflection on limitations and future directions for research in building more accessible and usable image exploration systems.

5.1 Mental Load of Exploration

When comparing a text-based system with touch-based systems, there is a tension between mental effort and image understanding. A text-based system requires less mental effort than a touch-based system because a text-based system is simpler to interact with and places less importance on users' abilities to recall the explored information. This is because a list of text can easily be read linearly using typical screen reader gestures, while a touch interface necessitates more careful interactions to traverse through all of the image elements. This difference led participants to eventually express frustration when using a touch-based system because they sometimes could not locate every element in an image, even after spending a lot of time with it. This mental effort could also potentially vary based on a participant's previous experience with tactile graphics, and future work could investigate this.

Additionally, when forgetting a piece of information (e.g., title of a book in image 2C), participants who were using a touch-based system often could not re-locate that information quickly, while those who were using a text-based system were able to with ease. On the other hand, when using a touch-based system, participants were able to explore an image more thoroughly, which encouraged skepticism towards incorrect image captions and allowed participants to provide more accurate and detailed explanations as to why the captions are correct or not. Notably, touch provides relative positions of objects, which empower blind users to form a mental model of an image.

Whether the value of image understanding triumphs the cost of mental effort depends on the amount of available time and importance of the images users want to explore. Participants said that they would opt to using a touch-based system whenever they have the time to thoroughly understand an image. This notion is best summarized by P6 who said *"I think it would be helpful to be able to have a quicker or speedy version where you have text... and then if you want to engage with it more on your own terms or have some time, you would use the touch approach."* Additionally, personal importance of an image seems to affect the willingness to spend additional time with that image. For instance, a picture taken together with family members hold much more value than an image of a park. Interestingly, both P1 and P8 who mentioned that they are cat owners spent more time exploring image 3B than the majority of the participants.

Ultimately, the decision to sacrifice mental effort for image understanding depends on the user and their situation; this hints at a system that allow its users to have the autonomy to choose between mental effort and understanding by adapting to the amount of information users want.

5.2 Practicality of Exploration

Our goal in this work was to better understand how image exploration could be used as a tool for finding errors in auto-generated captions. Towards this goal, we chose to evaluate three real-world systems, each with their own limitations. Although exploration systems generally presented more accurate information than natural language captions, they still occasionally presented incorrect or incomplete information. While participants were not informed of inaccuracies in the exploration systems so as not to introduce bias, these inaccuracies could still influence their ratings of each caption and their understanding of the images. Generally, participants' previously used intuition strategies for assessing accuracy (assessing caption grammar and using prior knowledge of reality) did not apply to the exploration systems. Future work could further investigate how people make judgements of correctness for similar pieces of information (i.e., if given two similar captions, which one do they trust more). Additionally, future work could research generally what information helps users best judge accuracy.

In the future, image exploration systems might also help create better captions. By leveraging exploration patterns, captions might be generated that contain information that users may deem more relevant. Can we leverage the interaction data (path and order, how long they dwell) to provide training data to make caption models better and more interactive? Furthermore, if blind users are providing their own images for exploration, can we leverage their contextual understanding of the image (such as capture time, location, intention, and camera framing) to enable blind photographers to generate and label their own image datasets to train AI-based systems, e.g., personal object recognizers?

5.3 Next Iteration of ImageExplorer

5.3.1 Combining Facebook and ImageExplorer Interfaces. Towards the end of the study, as an open-ended question we asked participants to design their ideal image exploration system. There was a consensus among 8 of the 12 participants that they want a system that could be flexible to the amount of information that they want at a given time. When they are not particularly interested in an image, they could ideally just read the high level information, and when they are really interested in understanding an image, they could receive much more details about it. Six participants further commented that they want a system that can toggle between Facebook's text-based summary and ImageExplorer: a text summary for quick exploration, layers for flexible information intake, and touch for spatial information when needed.

Such a system could first present information about an image in a text-based manner to enable quick and easy exploration of an image. The textual information could be grouped into different categories, much like how Facebook presents its results, though the exact categories could potentially be improved. Participants appreciated the "position information," "size information," and "elements by category" categories that Facebook provided. While they did not suggest changing the latter two categories, for "position information," they wanted the information to be grouped into more subcategories beyond just "left," "right," and "center." For instance, P11 wanted to know elements that are located in the upper right corner of their screen. Based on similar comments, we suggest presenting position information using a three-by-three grid. Additionally, participants wanted two more categories: "color" and "count." A "color" category could summarize the colors of each elements in an image, while a "count" category could summarize the total number of each elements in an image. While these two pieces of information were presented to participants as part of the other categories, participants wanted these to be organized separately since they were crucial to understanding an image further.

After briefly exploring an image for more information, if participants want to know the spatial information in an image, they should be able to switch to a touch-based interface, similar to that in ImageExplorer. The majority of participants stated that spatial information such as absolute and relative positions are essential to understanding an image and identifying errors in its caption. For example, it is much easier to find out that image 3B's caption is incorrect using touch than text because the object relationship is embedded in the spatial layout of the image. While a text-based interface is easier to use, a touch-based interface provides spatial information, which is difficult to convey through text.

Finally, participants wanted the touch-based interface to be multi-layered similar to ImageExplorer such that they can not only get a lot of information in an organized manner, but also explore the composition of complex objects and scenes. For instance, participants were able to confirm that image 2C contained a black book and not a box because its second layer information included text.

A key advantage of this system is that users would have the freedom to choose whether to access more information or not. Some may want to explore an image briefly with hopes to understand it at a basic level, while others may want to explore elements in detail to create a rich mental model of the image. Therefore, this next iteration of ImageExplorer is necessary to support different users and their use cases.

5.3.2 Providing Even More Information. ImageExplorer provided a a range of information about an image by strategically combining the results of multiple off-the-shelf deep learning models, which participants appreciated. However, every participant commented that they would prefer even more information to fully understand the images. We compiled commonly mentioned details that participants considered helpful or necessary in understanding images:

- Color (e.g., What is the color of the cat in image 2C? What is the color of the book in image 3A?)
- Size (e.g., What is the size of each of the dogs in image 1B?)
- Count (e.g., What is the exact number of dogs in image 1B? What is the exact number of zebras and giraffes in image 2B?)
- Action (e.g., Is motorcycle in image 1A being driven by the police officer, or is it parked? What are the people doing in image 1C?)
- Background Information (e.g., Are the animals in image 2B in a cage? What else is in the picture besides a bed in image 2A? What else is in the image besides a police officer and a motorcycle in image 1A?)
- Type (e.g., What kind of a cat is the cat in image 2C? What kinds of dogs are in image 1B?)

As mentioned in Section 4.6, the level of detail in a caption might affect its perceived accuracy. Therefore, providing more details by incorporating more context-aware and generalized vision-language models [55] and presenting them both textually and spatially could further encourage users' skepticism towards image captions.

6 CONCLUSION

In this work, we explored how various image exploration modalities could support BVI people in identifying errors in auto-generated captions. We presented a comparison of three image exploration systems, including *ImageExplorer*, a design probe that allows BVI users to gain multi-layered image information through touch. Our results indicate the usefulness of touch and layers in increasing image understanding, and demonstrate that additional quantities of information could be useful in enabling BVI users to assess errors in captions. Overall, ImageExplorer is a step towards understanding the role additional information plays in image interpretation, and for improving the design of future image understanding systems.

REFERENCES

[1] Apple. 2021. Human Interface Guidelines. https://developer.apple.com/design/human-interface-guidelines/

[2] Cynthia L. Bennett and Os Keyes. 2020. What is the Point of Fairness? Disability, AI and the Complexity of Justice. *SIGACCESS Access. Comput.* 125, Article 5 (mar 2020), 1 pages. https://doi.org/10.1145/3386296.3386301

[3] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 401–413. https://doi.org/10.1145/3461702.3462571

[4] 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

[5] 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

[6] Rajarshi Biswas, Michael Barz, and Daniel Sonntag. 2020. Towards Explanatory Interactive Image Captioning Using Top-Down and Bottom-Up Features, Beam Search and Re-ranking. *KI-Künstliche Intelligenz* 34, 4 (2020), 571–584.

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[8] Google Cloud. 2021. Cloud Vision API. https://cloud.google.com/vision.

[9] Microsoft Azure Cloud. 2021. Azure Computer Vision API. https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/.

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

[11] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

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

[13] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*. Springer, 15–29.

[14] Leah Findlater, Steven Goodman, Yuhang Zhao, Shiri Azenkot, and Margot Hanley. 2020. Fairness Issues in AI Systems That Augment Sensory Abilities. *SIGACCESS Access. Comput.* 125, Article 8 (mar 2020), 1 pages. https://doi.org/10.1145/3386296.3386304

[15] 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

[16] Cole Gleason, Amy Pavel, Himalini Gururaj, Kris Kitani, and Jeffrey Bigham. 2020. Making GIFs Accessible. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 24, 10 pages. https://doi.org/10.1145/3373625.3417027

[17] 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

[18] Jiangtao Gong, Wenyuan Yu, Long Ni, Yang Jiao, Ye Liu, Xiaolan Fu, and Yingqing Xu. 2020. "I Can't Name It, but I Can Perceive It" Conceptual and Operational Design of "Tactile Accuracy" Assisting Tactile Image Cognition. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 18, 12 pages. https://doi.org/10.1145/3373625.3417015

[19] Timo Götzelmann. 2016. LucentMaps: 3D Printed Audiovisual Tactile Maps for Blind and Visually Impaired People. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility* (Reno, Nevada, USA) *(ASSETS '16)*. Association for Computing Machinery, New York, NY, USA, 81–90. https://doi.org/10.1145/2982142.2982163

[20] 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

[21] Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, and Meredith Ringel Morris. 2020. Toward Fairness in AI for People with Disabilities: a Research Roadmap. *SIGACCESS Access. Comput.* 125, Article 2 (mar 2020), 1 pages. https://doi.org/10.1145/3386296.3386298

[22] Anhong Guo, Jeeeun Kim, Xiang 'Anthony' Chen, Tom Yeh, Scott E. Hudson, Jennifer Mankoff, and Jeffrey P. Bigham. 2017. Facade: Auto-Generating Tactile Interfaces to Appliances. 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, 5826–5838. https://doi.org/10.1145/3025453.3025845

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

[24] Seung-Ho Han, Min-Su Kwon, and Ho-Jin Choi. 2020. EXplainable AI (XAI) approach to image captioning. *The Journal of Engineering* 2020, 13 (2020), 589–594.

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. IEEE, New York, NY, USA, 2961–2969.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[27] Leona Holloway, Kim Marriott, Matthew Butler, and Samuel Reinders. 2019. 3D Printed Maps and Icons for Inclusion: Testing in the Wild by People Who Are Blind or Have Low Vision. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 183–195. https://doi.org/10.1145/3308561.3353790

[28] 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*. IEEE, New York, NY, USA, 4565–4574.

[29] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. 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, 1–14. https://doi.org/10.1145/3313831.3376219

[30] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123, 1 (2017), 32–73.

[31] Jaewook Lee, Yi-Hao Peng, Jaylin Herskovitz, and Anhong Guo. 2021. *Image Explorer: Multi-Layered Touch Exploration to Make Images Accessible.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3441852.3476548

[32] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[34] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*. 4768–4777.

[35] 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

[36] Microsoft. 2010. Windows Phone 7 Series UI Design & Interaction Guide. https://blogs.windows.com/windowsdeveloper/2010/03/18/windows-phone-7-series-ui-design-interaction-guide/

[37] Microsoft. 2021. Seeing AI. https://www.microsoft.com/en-us/ai/seeing-ai

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

[39] Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P. Bigham, and Shaun K. Kane. 2016. "With Most of It Being Pictures Now, I Rarely Use It": Understanding Twitter's Evolving Accessibility to Blind Users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5506–5516. https://doi.org/10.1145/2858036.2858116

[40] 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).

[41] Krishnan Ramnath, Simon Baker, Lucy Vanderwende, Motaz El-Saban, Sudipta N Sinha, Anitha Kannan, Noran Hassan, Michel Galley, Yi Yang, Deva Ramanan, et al. 2014. Autocaption: Automatic caption generation for personal photos. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 1050–1057.

[42] Ravi Rastogi, TV Dianne Pawluk, and Jessica Ketchum. 2013. Intuitive tactile zooming for graphics accessed by individuals who are blind and visually impaired. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 21, 4 (2013), 655–663.

[43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[44] Elliot Salisbury, Ece Kamar, and Meredith Morris. 2017. Toward scalable social alt text: Conversational crowdsourcing as a tool for refining vision-to-language technology for the blind. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 5.

[45] Fawaz Sammani and Luke Melas-Kyriazi. 2020. Show, edit and tell: A framework for editing image captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4808–4816.

[46] Lei Shi, Ross McLachlan, Yuhang Zhao, and Shiri Azenkot. 2016. Magic Touch: Interacting with 3D Printed Graphics. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility* (Reno, Nevada, USA) *(ASSETS '16)*. Association for Computing Machinery, New York, NY, USA, 329–330. https://doi.org/10.1145/2982142.2982153

[47] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[48] Abigale Stangl, Jeeeun Kim, and Tom Yeh. 2014. 3D Printed Tactile Picture Books for Children with Visual Impairments: A Design Probe. In *Proceedings of the 2014 Conference on Interaction Design and Children* (Aarhus, Denmark) *(IDC '14)*. Association for Computing Machinery, New York, NY, USA, 321–324. https://doi.org/10.1145/2593968.2610482

[49] 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

[50] Brandon Taylor, Anind Dey, Dan Siewiorek, and Asim Smailagic. 2016. Customizable 3D Printed Tactile Maps as Interactive Overlays. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility* (Reno, Nevada, USA) *(ASSETS '16)*. Association for Computing Machinery, New York, NY, USA, 71–79. https://doi.org/10.1145/2982142.2982167

[51] Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, and Chris Sienkiewicz. 2016. Rich image captioning in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. IEEE, New York, NY, USA, 49–56.

[52] Unsplash. 2021. Unsplash Image Dataset. https://unsplash.com/data.

[53] Violeta Voykinska, Shiri Azenkot, Shaomei Wu, and Gilly Leshed. 2016. How Blind People Interact with Visual Content on Social Networking Services. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1584–1595. https://doi.org/10.1145/2818048.2820013

[54] Hanzhang Wang, Hanli Wang, and Kaisheng Xu. 2019. Swell-and-Shrink: Decomposing Image Captioning by Transformation and Summarization.. In *IJCAI*. 5226–5232.

[55] Junbo Wang, Wei Wang, Liang Wang, Zhiyong Wang, David Dagan Feng, and Tieniu Tan. 2020. Learning visual relationship and context-aware attention for image captioning. *Pattern Recognition* 98 (2020), 107075.

[56] Shaomei Wu and Lada A. Adamic. 2014. Visually Impaired Users on an Online Social Network. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3133–3142. https://doi.org/10.1145/2556288.2557415

[57] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic Alt-Text: Computer-Generated Image Descriptions for Blind Users on a Social Network Service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW*

’17). Association for Computing Machinery, New York, NY, USA, 1180–1192. https://doi.org/10.1145/2998181.2998364

[58] Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. 2017. The Effect of Computer-Generated Descriptions on Photo-Sharing Experiences of People with Visual Impairments. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 121 (dec 2017), 22 pages. https://doi.org/10.1145/3134756

[59] Yu Zhong, Walter S. Lasecki, Erin Brady, and Jeffrey P. Bigham. 2015. RegionSpeak: Quick Comprehensive Spatial Descriptions of Complex Images for Blind Users. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI ’15)*. Association for Computing Machinery, New York, NY, USA, 2353–2362. https://doi.org/10.1145/2702123.2702437