What's in an ALT Tag? Exploring Caption Content Priorities through Collaborative Captioning

ANNIKA MUEHLBRADT and SHAUN K. KANE, University of Colorado Boulder

Evaluating the quality of accessible image captions with human raters is difficult, as it may be difficult for a visually impaired user to know how comprehensive a caption is, whereas a sighted assistant may not know what information a user will need from a caption. To explore how image captioners and caption consumers assess caption content, we conducted a series of collaborative captioning sessions in which six pairs, consisting of a blind person and their sighted partner, worked together to discuss, create, and evaluate image captions. By making captioning a collaborative task, we were able to observe captioning strategies, to elicit questions and answers about image captions, and to explore blind users' caption preferences. Our findings provide insight about the process of creating good captions and serve as a case study for cross-ability collaboration between blind and sighted people.

CCS Concepts: • **Human-centered computing → Empirical studies in accessibility**;

Additional Key Words and Phrases: Accessibility, blindness, captioning, collaboration

ACM Reference format:
Annika Muehlbradt and Shaun K. Kane. 2022. What's in an ALT Tag? Exploring Caption Content Priorities through Collaborative Captioning. *ACM Trans. Access. Comput.* 15, 1, Article 6 (February 2022), 32 pages.
https://doi.org/10.1145/3507659

1 INTRODUCTION

In recent years, visual communication has become an important element of online social interaction. With the rise of social media and online video, information is increasingly shared as a combination of textual data and images or video. Even platforms that were originally designed to share text, such as Twitter, increasingly use imagery as a way of communicating. For example, a recent study found that more than 40% of popular posts on Twitter now contain some kind of image, and that the posts of texts are often tied to understanding the image content [38].

One major challenge that arises as we move toward image-based communication modes is the challenge of accessibility for blind and visually impaired people. Text-to-speech engines, such as screen readers, are efficient for browsing text but cannot interpret multimedia content without a meaningful caption or an alt tag. Alt text compliance is improving; a study in 2006 found that

6

This work was supported by the National Science Foundation under grants IIS-1619384 and IIS-1652907. Any opinions, findings, conclusions, or recommendations expressed in this work are those of the authors and do not necessarily reflect those of the National Science Foundation.
Authors' address: A. Muehlbradt and S. K. Kane, University of Colorado Boulder, Department of Computer Science, 1111 Engineering Drive, 430 UCB, Boulder, CO 80309 USA; emails: annika.muehlbradt@colorado.edu, shaun.kane@colorado.edu.

1936-7228/2022/02-ART6 $15.00
https://doi.org/10.1145/3507659

about half of the images on the web contained alt text, and a study in 2018 found that about 72% of images on the web contained alt text [25]. However, many instances of alt text have no content or contain one-word labels (e.g., the filename of the image) [4, 8, 24, 30, 32, 36, 38, 40, 43, 52]. Even when the alt text contains a description, blind caption consumers often need more information to understand an image [25, 42]. This presents an accessibility problem for people with visual impairments who engage in online social communities such as Facebook, Twitter, and LinkedIn.

Researchers have looked at computer-generated captions [21, 50] and automatic tagging systems [25] as an efficient approach to create and assign alt text. However, performance of these systems is poor, as much of the information that blind users consider important information about an image (e.g., context, aesthetics, and emotional valence) cannot be generated by current AI systems [44]. Computer-generated labels and automatic tagging systems are also error prone and have been shown to mistakenly present captions that misrepresent an image [16, 44]. This can be problematic, as blind users place a lot of trust in captions, filling in details to resolve differences between an image's context and an incongruent caption [33]. For images to be accessible, an alt tag must exist and also be of good quality (e.g., understandable, appropriate, useful).

To compensate for some of the shortcomings of AI systems, researchers have explored the use of human-powered approaches like crowdsourcing and real-time human assistance to generate and evaluate image captions [1, 2, 8, 51]. Crowdsourced captions rely on human annotators to describe images and are often more accurate than computer-generated captions. Crowdsourced captions can also be generated in near-real time by connecting blind users with sighted volunteers (e.g., [8]), and some services employ trained professional agents to provide general visual assistance via remote video (e.g., [5]). Providing high-quality captions remains a challenge, as online workers and volunteers are not trained in captioning and do not caption images according to accessibility guidelines [35]. However, professional services are costly.

In general, it is difficult to know if a given caption is good because little research has explored what makes a good image caption from the perspective of a blind consumer of that caption. In addition, it is difficult to evaluate caption quality because it may be difficult for a visually impaired user to know how comprehensive a caption is or to choose between different captions without ground truth, and it may be difficult for a sighted captioner to know whether a caption is missing something important or presents the right level of detail. In this article, we explore collaborative captioning to understand how both blind caption consumers and sighted captioners assess caption content. By making captioning a collaborative process, we were able to observe captioning strategies, questions and answers about captioned images, alternative captions for the same content, and blind users' caption preferences.

To explore what makes a good caption, and to engage both producers and consumers of captions in a discussion of caption quality, we conducted a series of collaborative captioning sessions in which six pairs of blind and sighted partners worked together to discuss, create, and evaluate image captions. In-person conversations between blind and sighted participants allowed us to observe the refinement of caption text from OK to good, and to capture caption preferences and blind and sighted participants' perceptions about what content should be in an image caption. Specifically, we examined (1) how caption quality can be shaped by engaging blind caption consumers in Q&A about image content, (2) what makes a good image caption, and (3) what the contextual differences are in image captions across different domains (e.g., fashion, places) and identities of the individuals creating the captions. In this article, we also introduce a method for collaborative captioning, share insights about composing captions that future captioning systems can use to improve alt text, and present a set of images and their respective "good" image captions.

2 RELATED WORK

Here we summarize prior research on evaluating image descriptions and in creating image descriptions. As our research applies a collaborative approach to creating accessible images, we outline prior work on crowdsourced and collaborative captioning and collaborating across different abilities.

2.1 Captions on the Web

Concerns about the accessibility of online images have existed since the early days of the web. Early work focused on developing best practices for creating alternative text for images on the web. The Web Content Accessibility Guidelines (WCAG) specify that alt text must convey all essential content and must fulfill essentially the same function or purpose as the image [17]. Other guidelines arose from the limitations of early screen reader software and browsers. Some screen readers had obscure settings for character length in alt text, imposing a de facto character limit on image descriptions [45], and some browsers did not support the longdesc attribute [42]. Slatin [45] advised to limit alt text to 150 characters and to include as much information as possible in the fewest characters without sacrificing intelligibility. Scholars also suggested that alt text should flow well with the surrounding content as it is intended to be read with the main text, that it should be written in normal prose, and that abbreviations should be avoided so it is suitable for speech synthesis (i.e., screen readers) [19]. These early guidelines present general recommendations on how to insert alt text and what value to insert; however, they do not provide guidance about what should be described about an image and how to compose alt text.

2.2 Evaluating Caption Content and Quality

A number of researchers have explored ways of identifying important image information and how to transform it into comprehensible prose. Tang and Carter [47] offered a question-guided procedure for composing informative alt text that considers general information regarding who, what, when, where, and information regarding relationships (e.g., logical, spatial, and temporal). Nganji et al. [39] deconstructed these more general categories to define specific information that would be of interest to blind users including color, gender, season, location, message (i.e., what the purpose of this picture is), objects, origin, actions, number (i.e., how many there are of a specific item). Others considered stylistic properties of image descriptions such as the use of formal language (e.g., scientific names), type of writing (i.e., inform, persuade, entertain), and reader response rather than focusing exclusively on visual relevance [34]. These methods provide clear instructions to help guide the composition of generic alt text but may not be suitable for all image categories and types.

Recent research has emphasized that requirements for what makes a "good" image description vary significantly by the domain of the images and the context. Researchers have explored issues related to creating image descriptions for fashion [3, 15, 46], social media content [23, 50, 54, 55], and artistic images [12]. In addition to information about key visual elements, users want to know about the impression of clothes in fashion images [15], the photo quality, time and location a picture was taken when sharing photos [55], and the emotions and humorous effect of memes [23]. It is clear that user requirements for image descriptions vary on the information that is being described. The domain and context of the described information also may affect users' tolerance for misleading or incorrect information [44] and subjectivity [15]. In this study, we explore contextual differences in image captioning based on the domain, as well as upon the identities of the individuals creating the captions.

2.3 Crowdsourced Captioning

Researchers have explored human-powered services as a means to annotate images when content authors fail to provide alt text. Von Ahn et al. [1, 2] introduced crowdsourced captioning using the ESP Game and Phetch, two computer games designed to motivate users to create keyword labels and natural language descriptions for arbitrary images on the web. Both of these games support annotating large datasets of images but rely on sighted users to know what blind caption consumers will want to know about an image. Bigham et al. [7] proposed an alternative system, VizWiz, that helps blind users answer self-contained visual questions about images in real time. Blind users can ask specific questions about image content, and multiple answers can be pooled to provide a more comprehensive response. Still, crowd workers may not provide consistent levels of detail in responses and answers can be unpredictable. Similar systems address these challenges by engaging blind users and crowd workers in continuous interactions (e.g., Chorus:View [31]) to avoid minimalist answers, and by iterating on responses (e.g., RegionSpeak [58]) to get more reliable answers. These systems are expensive and require significant effort from blind users to engage in Q&A, and therefore they may not be suitable for captioning and describing images at scale. More generally, crowdsourcing may be unreliable because human services may not always be available, and crowd workers may misunderstand what they are asked to do, may be lazy or even malicious, and violate users' privacy by exposing sensitive information to strangers [9]. Morash et al. [35] explored how providing structured form inputs can improve the quality of crowd-generated captions.

"Friendsourcing," a type of crowdsourcing in which social network contacts perform tasks, has also been investigated as a method to caption images at low cost. Brady et al. [11] explored a support system that allowed blind users to ask their friends to help identify labels and read text. Blind users preferred not to use this tool to avoid causing extra effort or demanding attention from their social contacts. However, friend-sourced answers can often contain personal or contextual information [37] that can improve the quality of image descriptions.

Finally, crowdsourcing has been employed to improve machine-generated captions. In WebInSight, a web page is automatically annotated with alt text by retrieving HTML metadata to label image links, using optical character recognition to extract embedded text in images, and by allowing blind users to choose when to enlist human annotators [8]. More recently, Salisbury et al. [44] developed TweetTalk, a tool in which a computer vision system generates an initial caption and then supports a structured dialogue between a blind user and a crowd worker to enable the blind user to ask clarifying questions about the caption. The structured Q&A between blind and sighted users revealed that questions considered important by blind users include subjective issues (e.g., aesthetics, emotion), and researchers constructed a set of eight canonical questions most relevant to understanding social media imagery that human captioners and automated captioning systems can use to improve image descriptions. This suggests that dialogue between sighted and blind users can uncover users' preferences for caption content and assist in understanding how to refine captions from OK to good. Therefore, our work also takes a collaborative approach wherein blind and sighted users work together from the start to describe and caption images. We expand the prior work by investigating how the questions that users ask change depending on the context of the image and the identity of the users creating the captions. We also investigate how caption text is adapted for different audiences and different caption lengths.

2.4 Automated Captioning

Developing automated systems to generate captions may be the optimal solution for making media content accessible, and researchers have explored a range of different approaches. One approach is to derive alt text from pre-existing descriptions. Bigham et al. [8] proposed a system that generates

alt text from image context, metadata, and nearby text, and more recently, Guinness et al. [25] developed a system that uses reverse image search to find and apply existing captions to similar images across the web. For these systems to work, an image description must already exist, and there is no guarantee that the existing description is accurate.

A second approach is to retrieve similar captioned images from a database of human annotated images to generate descriptions. For example, Keysers et al. [27] developed a system to add labels to images based on previously stored, similar images, and Ordonez et al. [41] performed Flickr queries to collect 1 million image captions and then used this collection to associate images with existing image content and descriptions. Instead of copying captions directly from other images, Kuznetsova et al. [29] used computer vision to identify specific elements of an image and then generate descriptions by retrieving and combining relevant phrases from a database of pre-existing human annotations. This approach requires a curated, large dataset of human annotated images for improved accuracy.

A third approach is to utilize sentence templates that are filled based on the results of computer vision-based object detection. Object detections in an image are paired with words and then the words are placed into fixed sentence templates (e.g., [21, 28]). With this approach, captions can be generated without requiring related text or similar images with pre-existing descriptions. However, fixed templates are limited to one kind of utterance resulting in many similar descriptions for different images.

More recently, researchers have developed general-purpose image description systems that can provide complete descriptions of images from the real world. These approaches have been deployed to end users via systems such as Microsoft's CaptionBot [49] and integrated into social networking tools such as Facebook [55]. These tools have been demonstrated to provide acceptable captions for a range of images, but they may break down in facing some difficult images and may not fully capture the range of images that users may wish to know about. In particular, these systems do not work on images with specific stylistic choices such as action shots where the people and objects in the image are blurred. Resulting captions are also too general and do not capture users' specific preferences [57]. Although the automatic alt text provided by platforms like Facebook is always available, it is not yet trustworthy to blind users and of poor quality compared to alt text written by humans [22].

2.5 Cross-Ability Collaborative Work

Our work is an example of cross-ability collaborative work between a blind screen reader user and a sighted partner/sighted friend (SP). Researchers have explored how people with disabilities work together with others when conducting tasks at home [13], in the workplace [12], while shopping [56], when playing board games [26], to navigate [53], when programming [48], to create and edit documents [20], and when participating in online communities [14]. In these studies, sighted people often misunderstood the access needs of their collaborators and blind people had to educate their peers on how to best assist them. Bennett et al. [6] highlights that access does not result through the support of technology or from sighted collaborators, but through the work people with disabilities do to co-create accessibility. Systems and services aimed to assist blind people can benefit by not seeing blind people as passive recipients of assistance and, instead, recognizing their expertise in creating access [6, 59].

In this work, we consider how blind screen reader users not only can be passive caption consumers but also can become caption creators and editors. We consider how collaboration between blind and sighted captioners can improve captions and how the design of collaborative captioning tasks can impact the quality of the collaborative relationship between users.

Table 1. Participant Demographic Information

Participant	Age	Gender	Relationship	Nature of Blindness	Duration
VIP1	†	Male	Married	Blind w/ some light perception	Gradual vision loss since birth
SP1	†	Female			
VIP2	28	Male	Friends	Blind w/ some light perception	Since birth
SP2	39	Female			
VIP3	22	Male	Roommates	Total blindness	Since birth
SP3	20	Male			
VIP4	64	Male	Married	Total blindness	Gradual vision loss since birth
SP4	65	Female			
VIP5	26	Female	Friends	Total blindness	2 years
SP5	24	Female			
VIP6	38	Female	Friends	Total blindness	4 years
SP6	34	Female			

Note: Entries marked with a dagger (†) indicate that the participant chose not to answer that question.

3 COLLABORATIVE CAPTIONING STUDY

In this study, we created and deployed a collaborative captioning task to explore cross-ability collaboration between blind and sighted partners. In this task, participant pairs talked and worked together to create captions for photographs across three different categories: people, places, and events. The study was motivated by these goals:

(1) To gather information about the desirable attributes of different types of photo captions by documenting the collaborative creation of captions;
(2) To identify similarities and differences in desired captions between different groups and image types; and
(3) To explore the use of collaborative captioning tasks as a form of cross-ability collaboration.

3.1 Participants

We recruited six visually impaired participants (VIPs) (four male, two female) through campus email lists and word of mouth. The nature of visual impairment varied across participants, with three participants describing their impairment as total blindness from birth (or early age) and two describing gradual vision loss. One participant described her vision loss as sudden. All VIPs identified as being proficient screen reader users. All of our participants had earned a higher education degree, and all but two participants were native English speakers.

We also asked each VIP to invite their SP to the study, and they did so. Relationships between VIPs and SPs varied; three participant pairs described their relationship as friendship, two pairs were married, and one pair said they were roommates. We choose pairs of participants who knew each other to promote open and honest discussions between friends and partners and to reduce the time needed to orient the participants.

Participants' demographic information is summarized in Table 1.

In addition to demographic information, we asked participants about their experiences creating and editing alt text, social media use, and collaboration tools and practices. Only SP12 reported having experience creating alt text for images, although all of the sighted participants had some experience describing family photos, posters, or other visual media to their blind partner or friend. All participants reported visiting at least one social media platform a week, and many participants engaged in at least three online social communities a week. Most participants were also used to collaborating at least on a weekly basis and were familiar with many collaboration tools

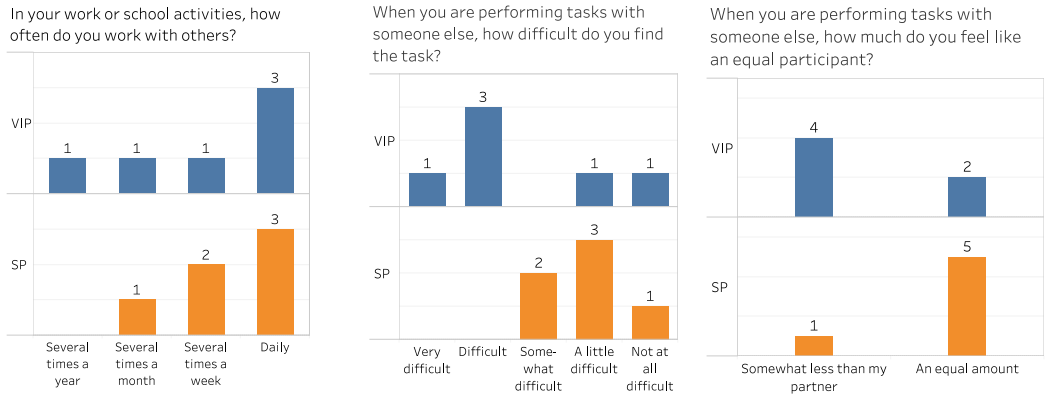

Fig. 1. Participants' responses to interview questions about collaboration. Participants reported collaborating with others at least on a weekly basis. VIPs found it more difficult to collaborate with others than SPs. VIPs felt less like an equal participant when collaborating, whereas SPs felt like an equal participant.

such as Google G-Suite, email, and various messaging applications. On average, blind participants found it much more difficult to collaborate with others than sighted participants. Similarly, most blind participants reported participating somewhat less than their partner, whereas most sighted participants reported participating an equal amount. Participants responses to questions about collaboration are summarized in Figure 1.

3.2 Study Design

The study took place over one session lasting approximately 120 minutes. In each session, blind and sighted participants worked in collaborative pairs. Each participant was given their own laptop for the study session. These laptops all featured Windows 10, Google Chrome, and the JAWS screen reader. VIPs were given the option to bring their own laptop to use their preferred device and software; four participants brought their own laptops. All tasks took place in a shared Google Docs document.

After participants completed the consent process, we then provided an introduction to the goals of the study, asked demographic questions (summarized in Section 3.1), and introduced participants to the laptops they would be using.

The main body of the study included two study tasks. In the first task, participants collaboratively created captions for five images in three categories (people, places, and events). In the second task, participants discussed how to shorten and simplify the captions they created. After these tasks, the research team interviewed the participants about their experience.

3.3 Image Set

All participants in the study created captions for the same set of 15 photographs. These photographs were based on images from social media, as these represent a significant accessibility challenge for people with vision impairments [55, 57], and because they would be of interest to the general population (as opposed to scientific diagrams, for example). We selected a set of photographs from the real world so that participants would need to choose which elements of the photo to describe and where to provide detail.

We selected images from three categories commonly found on social media: people, places, and events. For images of people, we selected images from a department store's Instagram feed wherein individuals are showing off an outfit. For images of places, we sampled photos from the Instagram hashtag "architecture," as these images featured unique visual elements such as irregular shapes

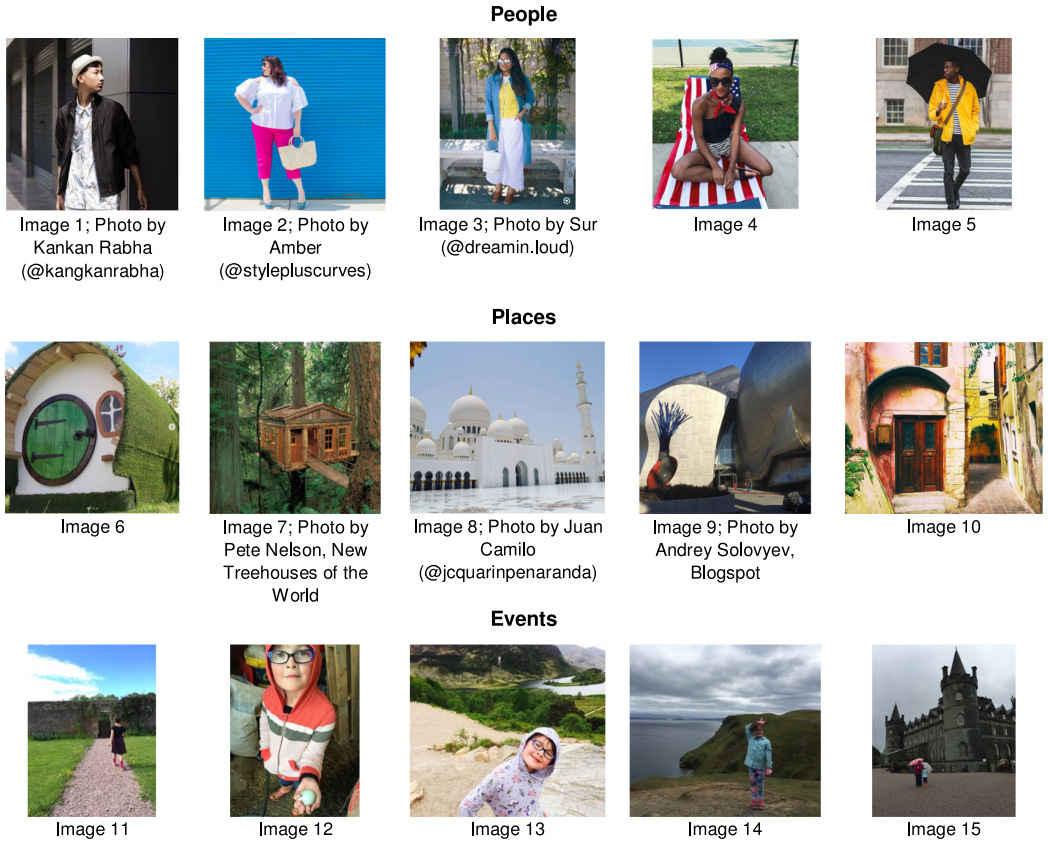

Fig. 2. The set of image participants captioned, depicting different people, places, and events.

and unusual photo angles that may be especially difficult for machine captioning systems to caption [57]. For events, we could not find an appropriate set of images on the web and thus used a set of travel photos from a personal contact (with permission). This category represents images commonly shared among family and friends. Although the visual elements in these pictures may be similar to the images in the "people" category, the context for these images is different. The images in the set are related and detail a story, and each includes a combination of people and scenery. The final set of images is shown in Figure 2.

For each study session, we copied all of the images into a single shared document on Google Docs. Both participants could access the document via their laptops. Participants were instructed to write their captions beneath the image. The order of the image sets and the order of images within the set were randomized for each study session.

3.4 Study Tasks

During the main part of the study, participants performed two tasks: creating image captions and shortening image captions. As the shortening task involved working with the output of the creating tasks, participants first created captions and then shortened them.

3.4.1 Task 1: Creating Captions. During this task, participants created image captions for the three sets of images. Participants were given 15 minutes to complete each set and were instructed to caption as many images as possible during that time, in whatever order they preferred. To ensure

that the ordering of image sets and images did not affect the caption content, we randomized the order of the images in each set for each participant pair and counterbalanced the order of the sets across participant pairs. Participants captioned a single set of images at a time and were not allowed to skip forward or go back to a previous image set.

As participants may not have previously completed this type of task, we suggested that the sighted participant first describe each image to the blind participant and that participants then work together to create a caption but instructed participants to choose a process that worked best for them. Each participant could write in the shared document.

One challenge in soliciting "good" captions for complex images is that the captioner must know the purpose of the image to know what parts of the image to emphasize. For example, consider a photograph of a family eating dinner at a table: if the image is part of a food blog, viewers of the image would likely want to know about the food itself, whereas if the image were shared on a social media, viewers might be more interested in the group of people at the table; knowing the context of use will help the captioner make decisions about what to prioritize and what to leave out. To provide some structure for the captioning activity, we presented each context with a particular task that the participants were meant to be performing when captioning the images:

People: "Your favorite clothing company is working on creating accessible images for a style catalog. Your task is to write captions for images in this style catalog."

Places: "Your blind friend is looking for inspiration for his new house. He is working with an architect to draft the designs. He asks you for help. Your first task is to write captions for these images."

Events: "You want to surprise your friend Ann with a captioned photo album of Ann's trip to Ireland with her daughter Betty. You have selected her five favorite photos from the trip. Your task is to write captions for these images."

3.4.2 Task 2: Shortening Captions. On the web, it is common to provide both a brief alt tag and an extended longdesc tag. To understand the process of creating both long and short captions, we asked participants to convert their original descriptions into shorter texts. We included this task partly as a way to identify which parts of the original captions were more important and which were less important, assuming that participants would include the most important terms in the shortened caption.

In this task, we instructed participants to review each of the captions that they had created in the first task and to shorten them. Participants were given 15 minutes to shorten all of their captions (across all image sets) and were instructed to shorten as many captions as possible. We asked participants to reduce the length of each caption by approximately half.

3.4.3 Post-Task Interview. Following completion of the two study tasks, we asked participants about their experience completing each of the study tasks. We asked participants to rate the difficulty of each task, to identify images that were most difficult to caption, to rate how effectively they contributed to the task of captioning images and shortening captions, and about how satisfied they were with the captions they created. Because we also wanted to solicit general feedback about the process of image captioning, image captions, and the study, we encouraged open discussions between participants. As some of the questions can be considered sensitive and we wanted to elicit honest answers, we let participants choose if they wanted to be asked these questions separately or together.

3.5 Data Collection and Analysis

We recorded audio and took detailed observational notes to capture the interactions between participants. We analyzed the data using an open coding approach [18] and conducted two stages

Table 2. Seven Interaction Types Identified in the Activity Recordings

Name	Verbally describing image content (Describing)	Physically describing image content (Gesturing)	Asking a question about image content (Asking)	Discussing preferences (Preferences)	Answering a question or providing a clarification (Answering)	Reading out loud and suggesting changes to caption text (Editing)	Coordination (Coordinating)
Description	Saying out loud features or context of image	Making gestures to illustrate objects	Verbally asking about objects, relations, context, etc., or requesting confirmation about an assumption	Any experience or preference about how to describe and caption images (blind participant only); talking about what a person would want to know or assessing caption text quality	Any response intended to answer a question (e.g., restating a description to make it more intelligible, affirming an idea, or directly answering a question) or correcting a statement	Saying out loud caption text or suggesting changes; saying out loud the intention of making a change or soliciting feedback or asking about how to phrase caption text (i.e., what do you want me to write?)	Saying out loud the intention to do something other than typing or asking about work assignment, assigning roles or ongoing tasks, or asking about progress
Example	SP1: "The ceiling of the house is a round shape and it's full of the grass and it's green."	SP1 touches VIP1's arm to show the length of the sleeve of a blouse as it is shown in image 7.	VIP5: "Oversized pants? Like is the waist wide?" VIP1: "Is it like located in a forest or something?"	VIP5: "So I would say let's describe people, face... maybe scenery, but first people, and faces, and what they are doing is important."	VIP3: "What color lipstick?" SP3: "Red. But not a... not a dark red. A light red."	SP2: "So what I've written is in a European alleyway is an entrance of a house consisting of oak, French-style doors..."	SP3: "Are you ready to move on to the next image?"

of analysis. During the first stage, we analyzed the conversations to collect types of interactions between participants. We developed a coding manual to describe the general interactions between the two groups; this was led by one of the project team members and supervised by the other. The interaction types are summarized in Table 2.

In the second phase, we collected all questions asked by one participant to their partner during the study. Our analysis of these questions is described in detail later.

To analyze the content of descriptions, we collected the final version of each caption from the shared document. We analyzed the frequency of parts of speech (e.g., noun, adjective, verb) and the frequency of common words using the natural language toolkit (NLTK) in Python [10]. We used the word frequency analyses to identify themes within and across different image categories (discussed in Section 4.1).

We also measured the similarity between caption texts using the cosine similarity measure. First, we converted the tokenized and lemmatized caption texts into vectors of term frequency (TF). We counted the number of times each term occurred in the document ("term frequency") and used the result as the term score in the vector. We did not weight the score of the words by their

relative frequency in the captions, as we were interested in the repetition of words. Repetition of a word in a caption implies increased importance of the word in understanding the content of the image. Finally, we computed the cosine similarity between the vectors to obtain pairwise similarity measures between each of the captions (discussed in Section 4.1).

3.6 Limitations of the Study Method

This study is among the first in creating and deploying a cross-ability experimental task in a lab setting. Perhaps unsurprisingly, conducting the study was often more difficult than for other lab studies, due to the challenge in coordinating with a person with a disability and their chosen partner, preparing multiple participants simultaneously, and navigating accessibility issues related to the technology. Furthermore, as this study design was fundamentally new, we made decisions based largely on our prior research experience and based on our piloting of this study protocol.

Given these challenges, we note some limitations of the current study. First, because recruiting and coordination was difficult, we were limited in the number of participants who could take part in the study. Second, because of the overhead in setting up the study and coordinating between multiple participants, we were limited in the categories of images discussed, the number of images captioned, and the amount of time given to caption each image. Finally, we did not collect detailed information about our participants' backgrounds, such as extended details about their prior experiences in cross-ability collaboration, the length and progression of their relationship, and their general language and communication abilities. Thus, we cannot report on the effects of prior experience, relationship status, or language ability on the production of the captions themselves.

Despite these limitations, we were able to collect a variety of data about how pairs of blind and sighted partners caption different types of images, and how they work together to do so. This data is presented in the following. We discuss opportunities to build on these findings in Section 6.

4 FINDINGS

In this section, we present an overview of the captions created by participants, their performance in creating captions, and their feedback on the process. Due to the small size of this study, we have omitted statistical tests for these data.

4.1 Caption Content

Participants generated a total of 72 captions. Not all participants were able to caption all 15 images in the allotted time. Two pairs captioned 9 images, one pair captioned 11 images, two pairs captioned 14 images, and one pair captioned all 15 images (avg. 12; min. 9; max. 15). All participants captioned at least two images from each of the three sets.

For the first study task, participants were instructed to create whatever captions they felt would be most appropriate for their assigned task. As such, the content of the captions varied widely. Example captions from each group and each image set are presented in Table 3.

4.1.1 Caption Length. The average caption length in words was 44.9 words (stdev. 79.0). Table 4 summarizes the length of captions written by different participant pairs in each category. The average caption length varied by participant pair, with some pairs generating captions that were much longer than others. There was little variation in caption length across categories, and most participants generated captions that were of similar length regardless of image type.

4.1.2 Parts of Speech. We analyzed the parts of speech present in the captions. On average, captions contained 32.4% nouns, 15.0% verbs, 16.1% adjectives, and 36.5% other. Table 5 summarizes the distribution of parts of speech. Overall, the distribution of nouns, verbs, and so on was consistent across captions.

Table 3. Example Captions from the Study

	People	Places	Events
Pair 1	"A young man wearing a yellow light raincoat with black casual pants and an umbrella. He is carrying a green sport bag." (Image 5)	"Image of a public place (maybe a museum). The architecture looks unusual. The material seems to be metal and the walls are not square shape. People are coming to the entrance way." (Image 9)	"Betty standing near chickens and holding an egg in her hand and looking straight to the camera with a series face." (Image 12)
Pair 2	"Male outfit. White button down tunic top patterned with brown palm trees. Matching the shirt is a waist length ¾ black jacket with buckles along the jacket bottom edge and buckles along the sleeves. A complimentary accessory is a white short brimmed, flat top grass hat." (Image 1)	"This unique house appears almost barrel in shape. The entrance of the home is along the flat side of the 'barrel.' The roof and the side of the house are covered in either turf or face grass. The entrance of the home is a circular door made of wood that has been painted green and the door opens via barn door hinges. To the right of the door is a small window similarly shaped to the house with a t shape within the window." (Image 6)	"Betty looks so cute giving herself a pair of pink bunny ears with her fuzzy warm mittens. On this cold day the sun is hidden by clouds, which explains Betty's warm light blue coat. Behind Betty is the beautiful cliffs and coast of Ireland. The cliffs are to Betty's left and the ocean is to her right." (Image 14)
Pair 3	"Young African American man holding a black umbrella with a yellow handle. A green shoulder pack with orange interior fabric, and a brown shoulder strap. A yellow rain jacket, with a black and white horizontal striped shirt. Black cargo pants, and black leather lace-up shoes." (Image 5)	"A White marble building. Rounded domes with gold spine on top. Three tier muslim prayer spinnerette. 8 archways on each side of the 3 main archways, the three arches in the middle are taller than the rest. The centermost one is also taller, giving the whole set of arches symmetry. Image depicts a courtyard that contains a polished stone floor. The stone floor appears to be made of the same white marble as the buildings, with a floral pattern of red, yellow, and green stone." (Image 8)	"Closeup shot of her, blue rectangular glasses, green eyes, a hoody. The cuffs and wasteband are red. Rubber rainboots with flowery Persian style pattern 4 colors. She wearing jeans. She is holding some kind of bird egg in her left hand, and her right hand is in the hoody pocket. There seems to be mud and hay, and a pallet on its side, with a bag that says 'animal feed' faintly visible. There are buckets and bags of farm supplies near her." (Image 12)
Pair 4	"A young woman with long black wavy hair wearing brown high heeled sandals, loose ¾ length white slacks and a bright yellow cardigan over a white blouse. Over this she is wearing a denim blue outer garment that has ¾ length sleeves and hangs to her knees. She is holding a white purse and has rounded mirrored sunglasses." (Image 3)	"This is a log cabin tree house elevated at least as high as it is tall. The entrance is reached by a rope and wood bridge. There is a small covered front porch, a peaked roof and many windows. There is a large tree growing through the middle of the structure." (Image 7)	"The sun is finally out and we see Betty from the back walking up a gravel path towards a gate and a brick wall." (Image 11)
Pair 5	"Female model, off the shoulder black sleeveless top, black and white striped very short shorts, red bandana tied around neck, American flag head band, head band matches towel on chair she is seated on." (Image 4)	"Antique building appearing to be set in an urban area, double wooden doors with windows containing slightly tinted glass panes, asymmetrical arched awning." (Image 10)	"Anne and Bettie standing close together wearing nearly matching pokadotted raincoats and umbrellas, standing in front of a grey castle. Bettie wears bright orange rain boots." (Image 15)
Pair 6	"Young adult female model of Indian descent, wearing billowy, boxy linen pants that hit her ankle. She is wearing a non-form fitting buttery colored cardigan over a white top. Over the cardigan she is wearing a lite denim jacket that is a little boxy, with the sleeves cuffed to ¾. She is wearing large round mirrored sunglasses and is holding a white purse in her right hand, that has large round handles and is made of white leather. Her shoes are a open toed slide made of a medium brown leather. The background of the image is a gazebo with lilac flowers, and the model is standing in front of a cement bench." (Image 3)	"Large public building constructed of Aluminum. The walls undulate, and budge out. You can see many layers of building with height and width changes. The roofline of the building cascades at the different bends in the material. The entrance is squished in between the aluminum walls. The entrance is made of glass. There are tall windows above it that are rectangular. They are lined by thin aluminum. There is a temporary tent in front of the doors." (Image 9)	"Betty is standing in the foreground of the photo on the edge of a cliff in a blue rain jacket, holding her hand above her head. On the right side of the image, behind Betty, the verdant green hills drop to sea at a 45 degree angle. On the left hand side of the image we see the sea. In the background of the image there is a horizon line, created by the distant coast and hills. It is a grey day and the clouds have a wave like texture." (Image 14)

Note: Each row shows sample captions in each category produced by one of our participant pair groups.

Table 4. Average Caption Length (in Words) by Participant Pair and Category

	People	Places	Events
Pair 1	23.2 (stdev. 38.9)	25.0 (44.8)	22.0 (28.9)
Pair 2	56.3 (21.1)	81.0 (146.4)	51.7 (14.6)
Pair 3	57.8 (47.7)	76.0 (51.4)	72.8 (49.2)
Pair 4	53.4 (54.7)	56.3 (87.3)	16.2 (43.1)
Pair 5	22.6 (24.4)	20.4 (51.1)	25.3 (49.3)
Pair 6	88.7 (25.7)	75.0 (58.9)	70.0 (121.5)
Total	46.7 (55.5)	49.6 (92.7)	39.0 (83.4)

Table 5. Parts of Speech Present in Captions, Averaged across Groups

	People	Places	Events
Nouns	33.8% (stdev. 7.9%)	30.1% (6.3%)	33.0% (7.2%)
Verbs	15.5% (2.5%)	14.7% (5.0%)	14.8% (6.5%)
Adjectives	21.7% (5.6%)	13.6% (4.9%)	12.1% (3.6%)
Other	29.0% (5.7%)	41.6% (5.6%)	40.1% (6.1%)

Table 6. Average Cosine Similarity Score for Caption Texts across Image Categories

	All Captions	People	Places	Events
Mean Similarity (stdev.)	0.08 (0.10)	0.34 (0.13)	0.25 (0.14)	0.21 (0.12)

Note: Caption texts about people are more similar than caption texts about places and events.

Table 7. Number of Unique Words Used in Captions for Each Category and Participant Pair

	People	Places	Events	All Categories
Pair 1	37	0	48	85
Pair 2	36	0	85	121
Pair 3	76	0	128	204
Pair 4	107	37	0	144
Pair 5	0	55	83	138
Pair 6	104	58	0	162
All Groups	360	150	344	854

4.1.3 Caption Similarity. We measured the similarity of captions within groups using cosine similarity, which accounts for the use of similar words and word order. Examining similarity within an image category tells us how much variation there is between captions in that set: sets with high similarity tend to use similar words, and sets with low similarity do not. Table 6 shows the similarity scores for each of the image categories. Perhaps unsurprisingly, similarity within a category is higher than similarity across all captions. However, all similarity scores are relatively low (less than 0.5), indicating that captions across participant pairs and images were composed of different words.

Another way to examine the similarity and diversity of captions is by measuring the number of words used to describe particular images and categories. Table 7 shows the total number of unique words used to describe images in each category.

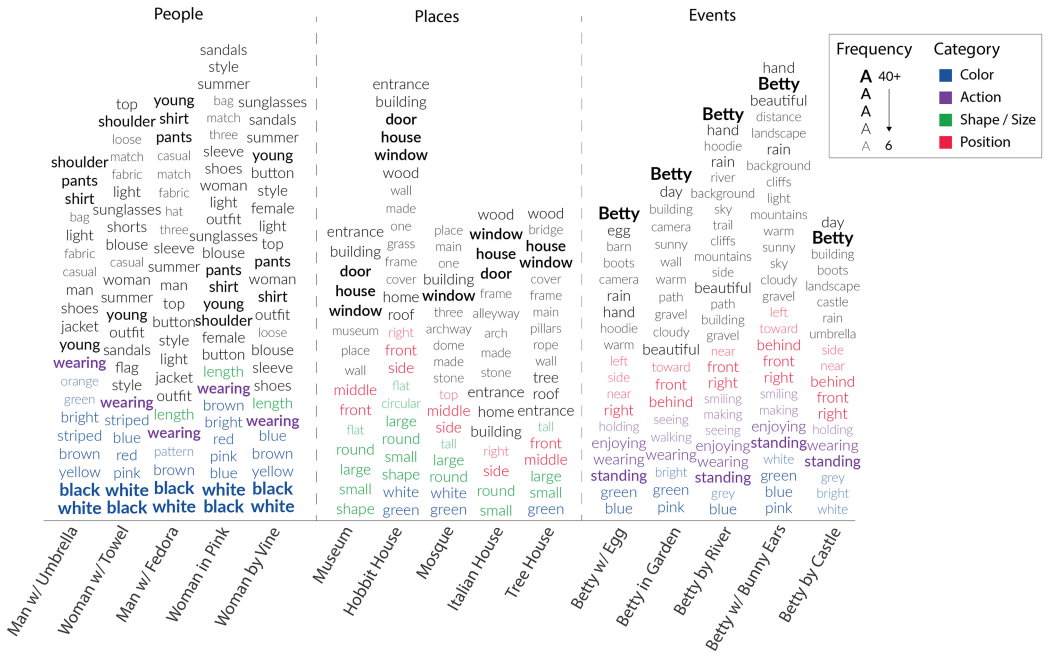

Fig. 3. The most common words that appeared in caption texts across image categories.

Table 8. Number of Times Colors, Shapes, Sizes, and Directional and Positional Words Were Used to Describe Objects in the Caption Text across Image Categories

	Colors	Shapes and Sizes	Directional and Positional Words
People	**126**	23	42
Places	54	**68**	**11**
Events	33	53	58

The number of unique words varied by category and by participant pairs. Participant pairs used fewer unique words to describe images of places compared to people and events. Three participant pairs did not use any unique words when describing images of places but used unique words to describe images in the other categories. Similarly, some pairs used more unique words, and perhaps had a greater vocabulary, than others.

We also examined words that were most common across all captions; these are illustrated in Figure 3 and Table 8. Perhaps unsurprisingly, the most commonly reused words typically involved descriptions of people (including clothes and body type), places (including buildings), and the characters included in the event images.

In captions about people, participants described things like clothing, accessories, hairstyles, gender, and age of the person in the image. The common words used were generic names of clothing items and accessories (e.g., "pants," "buttons," "sunglasses"), terms to describe gender (i.e., "female," "woman," "male," "man"), terms to describe age (i.e., "young"), and the color of clothes, hair, lipstick, and nail polish. Names of main colors were used more frequently than any other words and were mentioned 126 times.

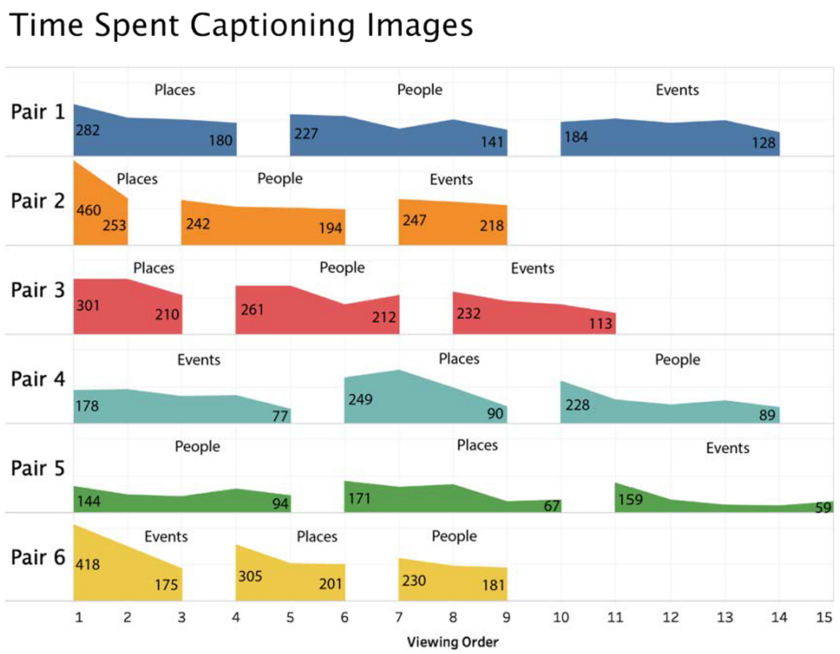

Fig. 4. The time participants spent generating initial caption text in seconds. Participants got faster at captioning images over time.

To describe places, participants detailed the main architectural components. The common words were terms to describe architectural structures (e.g., "window," "archway," "pillar"), materials (e.g., "wood," "stone"), and shapes and sizes (e.g., "enormous," "tall"). Terms to describe shapes and sizes were the most prevalent among common words. Surprisingly, participants used few directional and positional words such as next to, in front of, behind, and so on. In many cases, participants simply listed the architectural structures that were visible in the image and did not describe how structures were positioned or related to one another.

In captions about events, participants described the setting of the photo and the person in the photo, as well as mentioned ongoing activities seen in the photo. The common words used were directional and positional words to describe the composition of the photo (e.g., "behind," "front") and a handful of verbs to describe what a person is doing in a photo (e.g., "standing," "enjoying," "walking").

4.2 Captioning Process

We measured how much time participants spent captioning each image to see if participants got faster at creating image captions over time. We excluded off-topic discussion not directly pertaining to the image content and adjusted the timestamps if participants were interrupted (e.g., by their guide dogs). On average, participants spent roughly 180 seconds (3 minutes) captioning each image, although time spent on each image varied greatly (min. 42 seconds; max. 460 seconds; stdev. 90 seconds). We illustrate the time it took participants to caption images in Figure 4 and Table 9. In general, participants took more time at the start of each new category and gained speed as they became familiar with that category.

Table 9. Average Time Spent Captioning Each Category in Seconds

	People	Places	Events
Pair 1	186.0 (stdev. 38.9)	217.0 (44.8)	177.4 (28.9)
Pair 2	211.3 (21.1)	356.5 (146.4)	233.3 (14.6)
Pair 3	225.0 (47.7)	269.3 (51.4)	170.8 (49.2)
Pair 4	134.4 (54.7)	206.0 (87.3)	148.0 (43.1)
Pair 5	110.2 (24.4)	116.8 (51.1)	73.6 (49.3)
Pair 6	201.0 (25.7)	237.0 (58.9)	296.7 (121.5)
Total	173.1 (55.5)	214.7 (92.7)	170.7 (83.4)

4.3 Division of Work

We made image captioning a collaborative task to explore what information blind users want to know about an image and to capture blind and sighted participants perceptions about what content should be in a caption. To better understand how participants negotiated caption content, we measured participants' individual contributions to the captioning task. We also asked participants about how much they think they contributed to the task. Finally, we measured the perceived value of collaboration by asking participants if collaborating with their partner resulted in better or worse captions and if captions might have been different if they had completed the tasks alone.

4.3.1 Text Entered. We measured the division of work by counting the number of words that each participant generated by typing or dictating text to their partner. We counted dictating as a way to generate caption text because of the accessibility challenges of working in a shared document (further discussed in Section 5.2) and because SPs simply acted as human speech-to-text engines during these scenarios. Two VIPs opted to write and edit caption text using a screen reader, and four VIPs dictated their text to their partner. SPs generated 69.3% of the caption texts, whereas VIPs only generated 30.7% of texts.

SPs generated more text than VIPs. One reason this occurred is because SPs edited the caption text more frequently than VIPs. For example, SPs corrected mistakes such as spelling ("geans" to "jeans"), duplicate words ("the the"), and repetitive adjectives (e.g., black jacket with... the black jacket bottom edge"). The second reason is that VIPs sometimes relied on their partner's ability to skim the text and add or edit content quickly. After typing each caption text, VIP5 told her partner, "Now check my spelling!" Other times, VIPs asked their partner to add content, such as VIP3 asking his partner, "Add the sunglasses to that; those are important." It is likely that VIPs relied on SPs to edit caption text due to the accessibility challenges of editing a shared document (see Section 5.2).

4.3.2 Subjective Assessment of Contribution. We asked participants to gauge what percentage of the work they think they did in comparison to their partner, summarized in Figure 5. Two of the participant pairs felt that they and their partner did an equal amount of work. For three of the participant pairs, the SP felt that they did significantly more work than their partner and the blind participants felt that they did significantly less work than their partner. SPs and VIPs were in agreement about how much they contributed to captioning the images, and there was not an instance in which both the SP and VP claimed that they did more work than their partner. Participants' perceptions about how much they contributed aligned with our measurement of how much work each participant did (see the previous section).

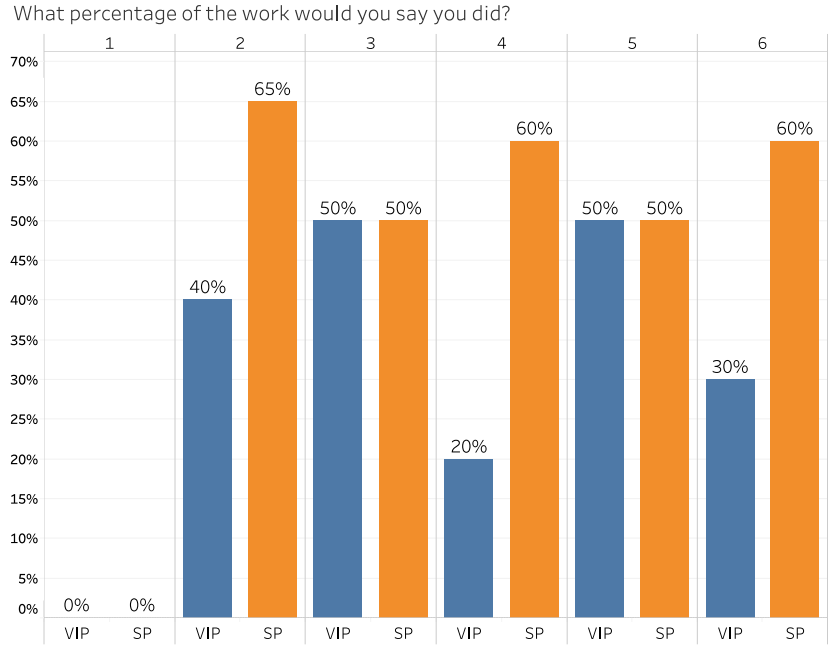

Fig. 5. The percentage of work that participants thought they did in the captioning task in comparison to their partner. Participant pair 1 did not answer this question.

We also asked participants if they quality of the long and short captions would have been better or worse without their partner. Five SPs believed the captions to be better with their partner, and only one SP believed the caption quality would have been the same with or without their partner. Three VIPs also felt that the captions they created with their partner were better than the captions they had encountered on websites and social media.

4.4 Image Description, Questions, and Clarification

We transcribed and analyzed participants' discussion about each image, including descriptions of the image and questions about the content.

4.4.1 Initial Image Descriptions. At the start of each captioning task, the SP would typically describe the image to the VIP. Although the final caption for each image was intended for a general audience, this initial description was intended to quickly introduce the image to the VIP. These initial descriptions were often personalized to the VIP, including references to shared knowledge, shared experiences, personal preferences, and previous images and captions, and were unique to participants' identities. Participants used these references to generated captions faster.

SPs often referenced famous buildings and scenery that they thought looked similar to the places depicted in the images. These references provided a quick but descriptive summary of an image's setting. When describing Image 8, three SPs mentioned that the building looked similar to the Taj Mahal, and for Image 9, three SPs mentioned that it reminded them of the Museum of Pop Culture in Seattle. For Image 6, two SPs mentioned that it reminded them of the Hobbit dwellings from the Lord of the Rings series.

Similarly, SPs recalled shared experiences to provide an explanation of the content shown in images. For example, to describe the shape of the blouse depicted in Image 2, SP1 recalled the blouse worn by a family friend at a wedding. SP1's partner understood what she was describing

right away. Participants also referenced their own clothing style or the clothing of their partner to describe items in Images 1 and 2.

Sometimes VIPs gleaned information about an image from their partners' reactions to and personal opinions about the aesthetics of image content. VIPs found their partners' personal opinions particularly useful in understanding image content because of their familiarity with their partners likes, dislikes, and quirks. For example, in the following conversation, VIP4 can glean what kind of clothing is depicted in Image 3 based on SP4's initial reaction:

> **SP4:** "Oh this is so ugly. This is not Eddie Bauer!"
> **VIP4:** "Oh it's not? So, it's like young people stuff?"
> **SP4:** "Yeah, it's all like, I don't know, it has weird proportions."

SPs were not shy about expressing their opinions, and interactions like these were common among participants, particularly for images about people.

Last, SPs made references to previous images to quickly establish the context of a photo and sometimes mentioned similarities and differences between images. For example, having previously explained Image 3, SP2 described Image 2 by noting "this also spring fashion," and after describing Image 13, SP1 explained Image 14 by first noting "this is Betty standing in nature again."

In general, VIPs found reference to shared knowledge, shared experiences, personal preferences, and previous images and captions to be useful. As VIP5 told her partner, "when you made that reference, I knew exactly what you were talking about, you can just skip the rest of the description."

To ensure that references were understood by their partner, SPs sometimes probed their partners' perceptions by asking questions about the reference. For example, when describing Image 1, SP6 asked VIP6 "when I say Miami, what do you think?" to confirm that her partner had the same mental visual of a typical Miami clothing style. Participants often probed for perceptions of size and shape to make certain that their description were not misleading. In several instances, SP5 asked her partner "if I say enormous, how big do you think?" Once participants established a reference, SPs used it over and over again to describe different images (e.g., "This is another enormous building.").

SPs sometimes relied on gesturing to describe the size and shape of objects that were not easy to express with words. SPs made on-body gestures to demonstrate things like clothing styles. For example, SP5 touched her partner's arm just above the elbow to demonstrate the length of the sleeve shown in Image 2. SPs also guided their partners' hands to demonstrate shapes of objects in the air. For example, to describe the round shape of the door in Image 6, SP1 took her partner's arm and moved it in a big circle to emulate the round door. Similarly, SP2 guided her partner's arm to his head and then moved it 2 inches away from his head to demonstrate the width of the brim of the hat in Image 1.

4.4.2 Clarification Questions. To understand what blind users want to know about different images and consider to be important in an image caption, we recorded VIPs' questions to their SPs. Because SPs did not describe everything in the image, nor did they know what their partner wanted to know about the image, VIPs asked general questions about image composition, aesthetics, and emotional valence, as well as specific questions about objects and people. Example questions are given in Table 10.

VIPs asked a total of 198 questions. The number of follow-up questions asked ranged from 0 to 9 questions per image, with an average number of 4 questions (stdev. 3) per image.

VIPs asked general questions to find out more about the setting of an image. About 18.3% of all questions were about image composition. Common questions included asking about the perspective of an image: "Is this like a close-up?" and "What is in the background?" VIPs also wanted to

Table 10. Examples of Questions Asked by Blind Participants during the Captioning Task

	Events	People	Places
Asking about an Attribute (e.g., shape, size, color) (36.3%)	"Is it a big egg or no? "Is it overcast or sunny in this one?"	"Short sleeves?" "What is the color of the hat?" "Is it tight?"	"Are they knobs, levers, or...?" "What color is the awning?"
Asking about an Action or Intention (10.2%)	"Is she laughing?" "Looking at what?" "Is she making a funny face?"	"Is she posing?"	–
Asking about Image Composition (18.3%)	"Which one? Left or right?" "What is in the background?" "What's she walking towards in this picture? "Do we see just her head?"	"Is this like a close-up?"	"Are there other houses next to it or just the one?"
Asking about Aesthetics (11.1%)	"Is it beautiful or no?"	"Is it like an ugly shirt or just … ?" "So, it's like young people stuff?" "Is it something you would wear?"	"Is it a nice place or like old?"
Asking about Emotions (9%)	"Is it a happy picture?"	–	"Oh, so you really like this place?"
Other (e.g., guessing (6.3%), missing detail (4.9%), did not hear partner (3.9%))	"Is she on the coast?" "What else is there in the image?"	"Like cobblestone or like gravel?" "A sombrero?"	"Covered with a grass roof or something?"

know how much of something was visible in an image. VIPs asked questions like "Do we see just her head?" and "Are there other houses next to it or just the one?" On occasion, VIPs wondered if the photo was staged or shot at a specific angle.

VIPs also asked their partner questions about aesthetics, which accounted for 11.1% of questions. Most often VIPs wanted to know if something was ugly or beautiful: "Is it beautiful or no?" and "Is this like an ugly shirt or just... ?" Other questions about aesthetics included asking "Is it a nice place or like old?" and "So, it's like young people stuff?"

Occasionally, VIPs asked their partner how they felt about an image. Questions about emotions accounted for 9% of the total questions and were the least frequently asked questions. VIPs asked questions like "Is it a happy picture?" and "Oh, so you really like this place?"

VIPs also asked specific questions about image content, such as asking about the color, size, and shape of objects or asking about the actions and intentions of people. Specific questions about objects were the most frequently asked follow-up questions and accounted for 36.3% of all questions. Questions about object attributes included asking "What is the color of the hat?" and "Is it a big egg or no?" Questions about people's actions and intentions accounted for 10.2% of all questions and included asking "Is she laughing?" and "Is she posing?"

Sometimes VIPs tried to guess the image content when their partner was struggling to describe an object (e.g., "Is she on the coast?"). Other times VIPs suggested terminology when their partner did not know what something was called (e.g., "Do you mean a Fedora?" and "Gauchos?"). Guesses and suggestions phrased as questions accounted for 6.3% of all questions.

Finally, VIPs asked questions to assure that their partner was not missing details that they felt would be integral to understanding the image content. VIPs asked questions like the following:

Table 11. Percentage of Stop Words, Nouns, Adjectives, and Verbs Dropped from the Initial Caption Texts by Image Category

	Total Words Dropped	Stop Words ("the")	Nouns	Adjectives (names of colors)	Verbs	Prepositions, Determiners, and Other
People	35.5%	37.3% (6.0%)	25.8%	25.4% (7.5%)	10.3%	1.2%
Places	40.2%	38.6% (4.7%)	26.6%	19.9% (2.1%)	9.1%	5.8%
Events	42.2%	39.3% (4.6%)	26.4%	18.5% (6.0%)	11.6%	4.2%
All Captions	39.2%	38.4% (5.0%)	26.3%	21.1% (4.6%)	10.0%	4.1%

"What else is there in the image?" "Is there anything else that stands out to you?" "Is there anything else important?" These questions accounted for 4.9% of all questions.

The remaining 3.9% of the questions did not relate to the content of the image. Sometimes VIPs did not understand what their partner had said and simply asked "what?" or "what did you say?" Other times VIPs asked how many images were left to caption or if their partner had finished typing.

4.5 Shortening Captions

After participants created captions for each of the 15 images (or as many as they could during the time limit), they then completed a second task that involved shortening those captions. To understand which parts of the original captions were most important and which were less important, we asked participants to convert their original captions into shorter texts. In general, participant pairs approached the task of shortening captions in a similar fashion. Both blind and sighted participants suggested changes, but only sighted participants edited the caption text in the Google Doc, as it was quicker for sighted participants to skim the text and locate the phrases that participants wished to change. First, participants tried to change the sentence structure of the caption without removing meaningful content. Next, participants removed redundant descriptions and portions of the caption that they felt was least important to understanding the image content. Not all participants chose to remove content from their original long caption, as some participants were able to sufficiently shorten the captions by changing the sentence structure. Some participants also chose to preserve the initial caption because they believed it could not be abridged in any way. Two participant pairs were unable to shorten all of their long captions in the allotted time. In the end, participants shortened 56 out of the 72 initial captions. The average length of the original and shortened captions are shown in Table 12.

Of all the words removed in the shortening process, approximately 38% were stop words such as "in," "to," "and," and "for." The article "the" was the most commonly dropped word and accounted for roughly 5% of all dropped words. About 26% of words dropped were nouns, 21% were adjectives, and 10% were verbs. Of the adjectives dropped, a little more than 4% were colors. Participants dropped fewer words (about 5% less) when shortening captions of people compared to captions of places (X^2 (1, $N = 1{,}775$) = 3.9731, $p = 0.04623$) and events (X^2 (1, $N = 1{,}222$) = 5.6419, $p = 0.01754$). There were no other statistically significant differences in words dropped by image category. The percentages of words dropped are summarized in Table 11.

Table 12. Average Length (in Words) of Original and Shortened Captions

	People	Places	Events
Pair 1 (original)	23.2 (stdev. 38.9)	25.0 (44.8)	22.0 (28.9)
(shortened)	17.0 (2.9)	14.4 (4.8)	16.8 (2.1)
Pair 2 (original)	56.3 (21.1)	81.0 (146.4)	51.7 (14.6)
(shortened)	30.8 (5.6)	35.0 (0.0)	†
Pair 3 (original)	57.8 (47.7)	76.0 (51.4)	72.8 (49.2)
(shortened)	†	41.0 (7.9)	†
Pair 4 (original)	53.4 (54.7)	56.3 (87.3)	16.2 (43.1)
(shortened)	30.3 (5.8)	25.4 (12.3)	8.0 (2.0)
Pair 5 (original)	22.6 (24.4)	20.4 (51.1)	25.3 (49.3)
(shortened)	8.5 (1.9)	6.4 (1.9)	8.2 (2.6)
Pair 6 (original)	88.7 (25.7)	75.0 (58.9)	70.0 (121.5)
(shortened)	50.5 (9.2)	42.0 (14.1)	38.0 (3.5)

Note: Entries marked with a dagger (†) indicate that the participant pair did not shorten any captions in that category.

4.5.1 Changes to Sentence Structure. To restructure sentences, participants replaced comparative clauses with compound words, removed stop words, combined sentences, and placed content in parentheses. Participants replaced long comparative clauses (e.g., appears as if, seems like) with shorter compound words formed by hyphenating nouns and adjectives (e.g., "museum-sized") and by fusing words together (e.g., "widelegged"). Because comparative clauses often contained qualifiers (e.g., appears, seems, suggests) that expressed doubt in the image content, removing comparative clauses and qualifiers made shortened captions seem more factual. In the following example, pair 5 replaced the comparative clause in the first sentence of the long caption with a compound noun in the short caption for Image 8. This shortens the sentence by six words and asserts that the building is Middle Eastern:

> **Long caption:** "This is an enormous white building that looks like it belongs in the Middle East. [...]"
> **Short caption:** "Enormous Middle-Eastern style white building with [...]."

When forming compound words, sighted participants did not always follow common convention (e.g., "Middle-Eastern" is not usually hyphenated) and sometimes used alternative punctuation such as slashes, ampersands, and plus signs (e.g., "black&white striped," "rope+plank bridge") without considering how these might be rendered by screen readers. In general, participants did not worry about using correct grammar or punctuation that could easily be parsed by a text-to-speech engine. Participants removed stop words such as pronouns (e.g., him, her), articles (e.g., the), position words (e.g., below, near), and transition words (e.g., in addition, similarly), and used short sentence fragments in place of complete sentences. Participants shortened captions further by combining multiple sentences into single run-on sentences containing comma-separated lists of adjectives, nouns, and phrases. In the following example, participant pair 2 shortened their original caption of Image 4 by removing stop words and replacing sentences with lists of short phrases:

> **Long caption:** "Young woman wearing a summer outfit that consists of a black sleeveless, one shoulder top that has a ruffled edge from the left shoulder to under the right arm. She is wearing black and white horizontal striped shorts. Complimenting the outfit is a red-white-and-blue flag headband and a red neck scarf."

> **Short caption:** "Female outfit, black sleeveless, one shoulder (left) top with ruffled edge from left shoulder to under right arm, and black/white horizontal striped shorts. Accessories: red-white-blue headband."

When abbreviating sentences, participants sometimes mistakenly altered the meaning of the text. Participants replaced phrases with compound words that are seemingly related but convey different meanings and mistook some words as synonyms that do not have the same meaning (e.g., "historical building" and "historic building"). To shorten the caption for Image 6, pair 3 replaced the description of the window and mullion with the specific architectural term "keyhole window," although a "keyhole window" is very different in shape from the window pair 3 initial described:

> **Long caption:** "[...] To the right of the door is a small window similarly shaped to the house with a t shape within the window."
> **Short caption:** "[...] To the right of the door is a small keyhole window."

4.5.2 Removing Content. Often participants were unable to shorten the long caption by only modifying the sentence structure and had to remove additional content to reduce it by half. Participants removed redundant adjectives and content that they felt was least important to understanding the image. Participants identified redundant adjectives as adjectives with similar meaning (e.g., "wooden timber framed house") and adjectives that were used as object labels and repeated whenever the object was mentioned (e.g., "[...] contains a polished stone floor. The stone floor appears to be..."). Participants also removed descriptions that they felt were implied by the context of the image and therefore superfluous. For example, when shortening the caption for Image 7, pair 1 excluded the description of the tree trunk in the tree house, as tree houses are often built around tree trunks:

> **Long caption:** "A wooden tree house in a very beautiful and dense forest with long green trees. The tree trunk is in the middle of the house giving a feeling of closeness to the nature."
> **Short caption:** "A tree house in a very beautiful and dense forest with long green trees that gives a feeling of closeness to nature."

Finally, participants removed content because they felt it was not essential to understanding the image. For captions about people, participants frequently removed descriptions of accessories like hats, sunglasses, and purses. For captions about events, participants focused on preserving explanations of what people are doing and general statements about where the event took place but removed detailed descriptions of the landscape. For captions about places, participants removed descriptions about architectural details such as the specific kinds of pillars, door handles, and awnings.

4.5.3 Similarity between Shortened Captions. We computed the cosine similarity score between shortened caption texts to examine if captions became more similar to one another as they were shortened. We looked at how similar captions are for a single image, for an image category, and for all images. A one-way repeated measures ANOVA was performed to compare the effect of shortening the captions on the cosine similarity score for each image, each image category, and across all images.

The average similarity score between all shortened captions was 0.11 with a standard deviation of 0.07. The average similarity was higher than the initial captions (avg. 0.08; stdev. 0.10). As with the initial captions, the similarity between shortened captions belonging to the same image category was greater than the similarity between captions as a whole. The average similarity score for shortened captions about people was 0.45 (stdev. 0.08) and significantly higher than the

Table 13. Average Cosine Similarity Score for Long and Shortened Caption Texts across Image Categories

	All Captions	People	Places	Events
Long Captions	0.08 (stdev. 0.10)	0.34 (0.13)	0.25 (0.14)	0.21 (0.12)
Short Captions	0.11 (0.07)	0.45 (0.08)	0.28 (0.11)	0.34 (0.14)

Note: Shortened caption texts about people are more similar than caption texts about places and events, and shortened caption texts are more similar than the initial captions.

initial captions in this category (avg. 0.34; stdev. 0.13; $F(1,20) = 22.32$, $p = 0.00013$). Shortened captions about events were also significantly more similar (avg. 0.34; stdev. 0.14) than the initial captions (avg. 0.21; stdev. 0.12; $F(1,22) = 17.833$, $p = 0.00035$). There was no statistically significant difference between the similarity of shortened and initial captions about places. The similarity scores are summarized in Table 13.

The similarity between captions for a specific image was also higher for shortened captions than for the initial captions. For Images 2, 3, and 4 (images about people) and Images 11, 12, and 15 (images about events), the shortened captions were significantly more similar to each other than the initial captions ($p < 0.05$). There was no statistically significant difference between similarity of shortened and initial captions for the other images.

The shortened captions for Image 12 were the most similar, with an average similarity score of 0.49 (stdev. 0.11). The shortened captions for this image were some of the shortest captions overall, and participants described the same visual elements and used the same or similar terms: "Betty," "egg," "hand," "holding," "holds." The high similarity can be explained by the image having relatively few distinct visual elements compared to other images as it is zoomed in. Image 15 also had a high similarity score for its shortened captions. This image also had few distinct visual elements, and participants all used the same terms to describe them: "castle," "umbrella," "jackets," and "rain boots."

There was one image for which the shortened captions had a lower similarity score compared to the initial captions—Image 4. For Image 4, participants did not always keep the same content when refining the initial caption. For example, some participants felt that the towel and headband were important, whereas others removed these two items from the description entirely and therefore also removed mentions of the American flag and the words "red," "matching," and "red-white-and-blue pattern."

The process of refining captions required participants to decide which parts of the caption text were most important. As the shortened caption texts are more similar to each other than the initial caption texts, this suggests that in some cases participants removed extraneous words while focusing on, and keeping, the most important ones.

4.6 Feedback about the Captioning Process

To understand the challenges of collaborative captioning, we asked participants a number of questions about the difficulty of the image captioning tasks. We asked participants to rate the difficulty of captioning images on a 5-point Likert scale (1 = Very easy, 5 = Very difficult), and we inquired about images and image categories that were particularly difficult to caption and about specific challenges of captioning images. We also questioned participants about the task of shortening captions and about their satisfaction with the long and shortened captions. Participants' responses to these questions are summarized in Table 14.

4.6.1 Feedback about Creating Captions. On average, SPs reported finding the task of captioning the images neither easy nor difficult (avg. = 2.83; stdev. = 0.55), whereas VIPs found the task to

Table 14. Participants' Responses to Questions about the Captioning Task: Which Images Were the Most Difficult to Caption? Which Images Were Easiest to Caption? What Was the Most Difficult Part of Captioning Images?

Participant	Most Difficult Image(s) to Caption	Easiest Image(s) to Caption	Difficult Aspects of Captioning Images
VIP1	People images	Places images	"Choosing phrases that other people could understand." "One's own knowledge about what's in the image makes the quality of the caption vary from person to person."
SP1	People images	Places images	"What is the focus of the image?"
VIP2	Places images	People images	"It's not just about what's in the image but also about the setting."
SP2	Places images	People images	"Getting a good idea of the image and then describing it succinctly."
VIP3	People images	Places images	"You had to simplify what they describe."
SP3	People images	Places images	"It's difficult to simplify the language." "People images were dang difficult for me cuz my care factor about clothing is zero and my knowledge about fashion is zero."
VIP4	Events images	Places images	"Knowing your audience's expectations. You don't know exactly what the other person is looking for."
SP4	People images	Image 7; Events images	"You got to get to the core of it and so when I was describing things, I wasn't putting in all the extra adjectives because you don't want all that information. But if it's in writing and it's for the general public then you're trying to be more descriptive [...] I'd find myself adding extra adjectives that I hadn't bothered to describe."
VIP5	Places images	Events images	"Architectural images were difficult because they were complex. I personally don't know what all of those things are called."
SP5	Places images; especially sculptures	Events images	"Deciding what was important [in the images]."
VIP6	Places images	Events images	"The easiest things to cut out [from the caption] are the things that make it human. But those things really matter."
SP6	Places images	—	"We have to be mindful of how we label pictures. If people haven't been [to Miami], we have to be mindful of how we label that." "I didn't have the right vocabulary to describe them."

be more difficult (avg. = 3.5; stdev. = 0.75). Only one SP found the task difficult, three SPs found the task neither easy nor difficult, and two SPs found the task easy. In contrast, three VIPs found the task difficult and three VIPs found the task neither easy nor difficult. No VIPs found the task easy.

Both VIPs and SPs reported that captioning some image categories was more difficult for them than others. Some participants preferred to caption images that they found personally interesting and reported having difficulty captioning images that they did not know much about or did not know the subject-specific vocabulary. Five participants (two VIPs and three SPs) reported that images about people were the most difficult to caption. SP3 explained that "[p]eople images were dang difficult for me [because] my care factor about clothing is zero and my knowledge about fashion is zero." Six participants (three VIPs and three SPs) found the images about places most difficult to caption because they felt that they did not know the proper terms for describing architecture. VIP5 noted, "I personally don't know what all of those things are called." Only one participant (a VIP) reported finding images about events the most difficult to caption.

Other challenges participants experienced were difficulties simplifying the descriptions or choosing appropriate language for the general public and deciding what things other people would find noteworthy in an image. Six participants (three VIPs and three SPs) reported having difficulties phrasing captions in a way that they felt other people could understand them. Some participants noted that they had trouble deciding what information is considered common knowledge and had difficulties identifying appropriate labels for places, people, and objects. As SP6 explained, "[w]e have to be mindful of how we label pictures. If people haven't been [to Miami], we have to be mindful of how we label that." At least five participants (two VIPs and three SPs) also felt that determining what is important in an image was challenging. It was difficult for both SPs and VIPs to know what information other caption consumers wanted to know about an image, or as VIP4 put it, "[y]ou don't know exactly what the other person is looking for."

There is no clear correlation between how difficult participants felt an image category was to caption and how much time they spent captioning images in that category. Participant pairs 2 and 5 spent more time captioning images in the category rated most difficult and less time captioning images in the category rated easiest. Participant pairs 3 and 6 also spent more time captioning images in the category rated most difficult but spent the least amount of time captioning images in the category rated as medium difficult. However, participant pair 1 spent more time captioning images in the category rated easiest and less time captioning images in the category rated most difficult. Finally, participant pair 4 did not agree on which category was the most difficult and the easiest.

4.6.2 Feedback about Shortening Captions. Shortening the captions was challenging as well. When we talked to participants about how difficult shortening was compared to creating the initial captions, seven participants (three VIPs and four SPs) reported that shortening the captions was more difficult than writing the initial captions. One reason was that participants did not know what content to remove from the initial captions. VIPs also reported that having to rely on their partner to make changes to the captions made it more difficult to shorten them (see Section 5.2).

Participants were also worried that shortening the captions might affect the quality. Some SPs and VIPs initially opposed to changing the long captions and shortened captions only after giving careful consideration to how removing elements from the text would affect its meaning. When shortening the captions, participants preferred to restructure sentences rather than to remove content entirely to preserve the descriptiveness of the caption text. In general, SPs were more reluctant to remove content than VIPs, and at least two SPs stated that all initial content of their captions was integral to understanding the images.

Although most SPs reported that the shorten captions were not as good as the long captions, some VIPs viewed the short and long captions as complementary where the short text represented

an alt tag and the long text represented a caption. In the words of VIP4, "[the short captions] are kinda like an alt tag that you'd get for images on the web."

5 DISCUSSION

In this section, we report on the success of the collaborative captioning activity, discuss strategies and obstacles that appeared during the process, and reflect upon the usefulness of collaboratively generated captions.

5.1 Captioning Strategies

To the best of our knowledge, this article reports the first instance of a collocated collaborative captioning activity between blind and sighted partners. Each participant group was able to caption most of the images in the image set in the allotted time, and all blind participants reported that they contributed to the task (although some reported that their contribution was less than that of their SP).

Although the task of captioning images in this specific manner was new to our participants, the participant pairs often developed strategies for more efficiently creating captions as they went along. Perhaps the most commonly adopted strategy was to build up a series of *cross references* between different images and captions, using previous images as points of comparison for the current image. Furthermore, because participants knew each other outside of the study context, they were sometimes able to reference shared knowledge as shorthand, replacing a lengthy explanation with a familiar comparison. This strategy highlights the value of shared context in communicating alternative descriptions and suggests that future image captioning systems may wish to draw upon the activities and background knowledge of the user reading the captions.

A second strategy that we observed over time was that after describing an image, SPs often reacted to the VIP's questions about that particular image by incorporating similar information into their descriptions of subsequent images. In this way, questions from the VIP served as *implicit feedback* about the desired content of a description. For example, after SP2 initially described a person's clothing (but not the color) in one of the images about people, VIP2 asked about the clothing's color. After this interaction, SP2 began to include color words in her descriptions of the following images.

This implicit feedback also guided the SPs to omit information that the visually impaired partner did not seem interested in. For pair 4, after starting on a new image, SP4 would first describe the image in vague terms. After some discussion, VIP4 would dictate the final caption for that image. While captioning the events image set, SP4 initially described materials and textures of the photo subjects' clothing; however, when dictating the final captions, VIP4 omitted those details. SP4 then began to omit these terms from her initial descriptions.

This use of implicit feedback further highlights the importance and usefulness of shared context: even among people who know each other well, a blind person's preferences for a particular type of description might be unstated or unknown but can be revealed through collaborative captioning.

5.2 Captioning as Cross-Ability Collaboration

Although many of our findings in this work specifically address image captions, a secondary goal of this research was to implement an example of a cross-ability collaborative task. In prior work, blind screen reader users reported that they faced difficulties in performing synchronous collaborative computing tasks with their SPs and as a result would often avoid these types of tasks, instead doing their work on their own time and alone [12]. This work thus serves as a rare example of a collocated, synchronous computing task between a blind screen reader user and an SP.

Due in part to the rarity of this type of collaborative task in the wild, we constructed a specific task for this study. In doing so, we attempted to create a task in which (1) both participants were able to make meaningful contributions, (2) the task leveraged the unique skills of both participants and thus required collaboration (without limiting our participant pool by requiring that participants have particular expert skills, e.g., audio production), and (3) the goal of the task served a real-world purpose. We made an effort to structure the task in a way that provided each participant with a clear role and motivation for completing the work. For example, when introducing and discussing the task, we emphasized our belief that the VIP served as an expert regarding the creation and evaluation of "good" captions.

As we designed this study with cross-ability collaboration in mind, we can also judge how effective we were in creating a fair and accessible task. Feedback from our participants was mixed: four of our six blind participants described the task as "difficult" or "very difficult," and blind participants rated the amount of their contribution to the task between 20% and 60%. Although we are confident that our study task met our own criteria for an effective cross-ability collaboration task, our design of the study task clearly influenced how each person took part in the activity and may have encouraged or inhibited certain types of behavior. As an example, by providing only an image, the VIP was reliant on the sighted participant to begin conversation of that image; providing both an image and a caption (perhaps a "bad" caption) might have resulted in a different division of labor. Clearly, more work is needed to understand the possible design space of these cross-ability tasks.

In observing study participants, we noticed a number of accessibility problems related to the cross-ability nature of this task. In particular, although we provided both participants with a laptop and access to the shared document, blind participants often experienced difficulty in editing the document while their SP was also editing. Specifically, participants experienced problems with cross-talk between their partner and their screen reader, struggled with items that moved around in the document as their partner edited it, and could be overloaded when their screen reader announced every change made by their partner. Over the course of the study task, blind participants tended to type less and to instead provide more structured commands to their partner, instructing them to navigate to specific areas or type specific text; two VIPs noted that instructing their partner was more efficient than typing themselves. These previously known accessibility problems remain an ongoing challenge for blind users, and a barrier to more equal collaboration between blind and sighted peers.

Sighted participants also experienced challenges related to cross-ability collaboration, specifically in balancing questions and suggestions from their partners with their own reading, speaking, and writing. For example, when writing out the "final" caption, SPs often asked their partner to hold their questions and comments, or to repeat a comment that they had missed. Although these problems further exemplify the challenges of cross-ability collaboration in general, they also diminish the role of the VIP even more, as after giving up on directly editing the document, they then found themselves sitting idle until their SP finished their work. These cross-ability collaboration problems affect participants of all abilities, although they seem to have an especially punitive effect on blind participants.

5.3 Are Collaboratively Generated Captions Good?

Although this article has focused on the activity of collaboratively generating image captions, we cannot yet conclude that captions generated in this way are "better" than captions generated through other means. Our purpose in conducting this research was not to create the best possible captions but instead to introduce a new process for creating captions and to study how blind and sighted partners can work together in creating captions.

However, even if we cannot objectively declare collaboratively generated captions to be better than others, it seems likely that captions created using these methods represent high-quality captions and thus may help us understand how to make captions better more generally. There are several reasons to believe these collaboratively generated captions serve as high-quality examples. First, these captions were co-created by sighted captioners who had a personal relationship with the caption consumer. SPs often commented on the knowledge or preferences of their partner and included shared knowledge within their descriptions. Second, these captions were clarified and refined based on follow-up questions from the caption consumer. In most situations, a caption writer must take their best guess about what the reader will want to know and may err on the side of information overload by providing all of the information a reader may ever want. Conversely, the captions generated here allowed for follow-up questions and iterative improvements. Finally, as discussed earlier, we found that partners changed how they described images as the study progressed, learning their partner's preferences from both explicit feedback and implicit cues.

Even if collaborative captioning is not feasible for all captions, creating and studying these captions may help us understand user needs and preferences related to image captions. Of course, excluding particular images such as charts and information graphics that can be directly and completely translated into text (e.g., [35]), caption quality may be subjective and may depend heavily on the particular context and preferences of the caption consumer.

5.4 Limitations of Our Collaborative Captioning Study

Captioning strategies changed over time as people learned their partners' preferences and developed common references. This suggests that there is a learning curve for this method, and that captions early in the process may be created more slowly (or less effectively) than later captions. However, our study design did not allow us to capture all differences in the captioning process between images and image sets. We did not compare how fast people captioned specific images. This made it difficult to determine whether participants were held up by specific images that may have been particularly difficult to caption.

As our study design focused on the process of creating captions collaboratively, we did not have an obvious method for measuring caption quality. Although it would be possible to have external raters judge the captions, it is unclear whether different raters in a different context would be the best judges of the caption quality, and we know of no standard, reliable method for evaluating caption quality. In the future, we could ask participants to revisit their captions later and reflect on their quality. The current study identifies commonalities in the decisions that participants made while captioning, which begins to explain what the captioners believe is most important about their captions.

Participants in the study chose their own partners. We chose this approach to increase the comfort of both participants, and to enable them to begin captioning with relatively little setup time. By working with a known partner, participants had the opportunity to refer to their shared knowledge and experiences. It is likely that the process of captioning would differ as the blind participants were paired with different partners, and it is possible that caption quality would differ between familiar and unfamiliar partners. The relationship between partner familiarity and the captioning process could be explored in future work.

6 FUTURE WORK

6.1 Future Collaborative Captioning Studies

In this article, we introduced a caption creation method based on cross-ability collaboration and presented an initial study of how blind and sighted partners work together to create captions. We are eager to continue this exploration of collaborative captioning in several ways.

First, we are interested in continuing our studies of collaborative captioning and extending this method to other user groups and media types. For example, we could explore how pairs of sighted people generate captions vs. a single sighted author, or pair a screen reader user with a previously unknown partner. We could additionally explore using this method to label photographs of everyday objects, images related to navigation (e.g., crosswalks and building signs), charts and information graphics, animation, or video. In collecting more examples of collaboratively generated captions, as well as the questions and clarifications involved in their creation, we could produce a dataset of collaboratively generated captions that may be useful to researchers studying image captioning and natural language interaction.

Second, as noted previously, our current study does not address potential objective or quantified measures of caption quality. Future studies could explore how different captioning teams and processes may lead to higher-quality captions. For example, we may study whether collaboratively generated captions compare to captions made by a single author or to those considered "best" online, how an individual's language and communication skills relate to caption quality, and whether previous experience with captioning or cross-ability collaboration leads to better captions.

Third, the design of the study tasks may have influenced the work done by each of the participants and may have guided them to participate in specific ways. In the future, we could explore ways to improve our method. We may choose to give blind participants more control in how the caption is created. For example, we could limit the sighted participant's role to describing the image and require the blind participants to write the caption. Alternately, both sighted and blind participants could use voice control to dictate, review, and edit the caption, and we could explore if explicitly separating the task into phases would lead to a more equal contribution between partners.

Finally, one theme that appeared throughout this study was the impact of an individual's prior knowledge, preferences, task, and current context on their evaluation of a caption. We may explore how this information could be represented in a personal profile, and how such a profile could be used to customize captions. This profile might contain information about other images that the user has encountered, background knowledge (e.g., does this user know specific terminology about architecture?), and style preferences (e.g., does this user frequently ask about people's clothing?). This information could be used to process and customize existing captions, or it could be used as an input to multimodal, spatial, and other new forms of captions [36].

6.2 Using Collaborative Captioning to Improve Image Captions Elsewhere

Although our present research has focused on documenting the process of collaborative captioning, we envision several ways in which the collaborative captioning process could be used to improve image captions. Most directly, future systems could include a synchronous or asynchronous collaborative captioning mode, in which a user browsing an image could request collaborative captioning help from a friend, a professional captioner, or perhaps even the creator of the original image. In this study, we established our blind participants as experts on what makes a good caption; in the future, blind captioners could share their expertise with others.

In addition to incorporating collaborative captioning into systems, it may be possible to leverage insights from the collaborative captioning process to improve automated and semi-automated captioning systems. For example, participants' follow-up questions represented information that was considered important and relevant to the image category. In the future, authoring tools could present these questions (customized to the image category) to a user who is attempting to caption an image, thus ensuring that this important information is not overlooked by the captioner.

7 CONCLUSION

Creating and evaluating captions for accessibility is a complex task, and it is also a task that fundamentally interpersonal. In most cases, the consumer of an image caption must trust that the caption writer faithfully described the image and must hope that the caption writer included all information relevant to their current task. Although image captioning relies upon this relationship between writer and reader, the relationship between these roles is typically indirect and obscured. In this study, we explored image captioning as a collaborative process, allowing for blind and sighted collaborators to discuss, question, and refine a caption. This study identifies how captions may be constructed collaboratively, how these captions may differ across individuals, and how collaborative captioning may serve as an example of cross-ability collaboration.

REFERENCES

[1] Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 319–326.

[2] Luis Von Ahn, Shiry Ginosar, Mihir Kedia, Ruoran Liu, and Manuel Blum. 2006. Improving accessibility of the web with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 79–82.

[3] Luis Von Ahn, Mihir Kedia, and Manuel Blum. 2006. Verbosity: A game for collecting common-sense facts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 75–78.

[4] Chieko Asakawa. 2005. What's the web like if you can't see it? In *Proceedings of the 2005 International Cross-Disciplinary Workshop on Web Accessibility (W4A'05)*. 1–8. https://doi.org/10.1145/1061811.1061813

[5] Mauro Avila, Katrin Wolf, Anke Brock, and Niels Henze. 2016. Remote assistance for blind users in daily life: A survey about be my eyes. In *Proceedings of the 9th ACM International Conference on Pervasive Technologies Related to Assistive Environments (PETRA'16)*. Article 85. https://doi.org/10.1145/2910674.2935839

[6] Cynthia L. Bennett, Erin Brady, and Stacy M. Branham. 2018. Interdependence as a frame for assistive technology research and design. In *Proceedings of the 20th International Acm SIGACCESS Conference on Computers and Accessibility*. 161–173.

[7] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, et al. 2010. VizWiz: Nearly real-time answers to visual questions. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*. 333–342.

[8] Jeffrey P. Bigham, Ryan S. Kaminsky, Richard E. Ladner, Oscar M. Danielsson, and Gordon L. Hempton. 2006. WebInSight: Making web images accessible. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*. 181–188.

[9] Jeffrey P. Bigham, Richard E. Ladner, and Yevgen Borodin. 2011. The design of human-powered access technology. In *Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*. 3–10.

[10] Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media.

[11] Erin Brady, Meredith Ringel Morris, and Jeffrey P. Bigham. 2014. Friendsourcing for the greater good: Perceptions of social microvolunteering. In *Proceedings of the 2nd AAAI Conference on Human Computation and Crowdsourcing*.

[12] Stacy M. Branham and Shaun K. Kane. 2015. The invisible work of accessibility: How blind employees manage accessibility in mixed-ability workplaces. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers and Accessibility*. 163–171.

[13] Stacy M. Branham and Shaun K. Kane. 2015. Collaborative accessibility: How blind and sighted companions co-create accessible home spaces. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2373–2382.

[14] Erin Buehler, Stacy Branham, Abdullah Ali, Jeremy J. Chang, Megan Kelly Hofmann, Amy Hurst, and Shaun K. Kane. 2015. Sharing is caring: Assistive technology designs on Thingiverse. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 525–534.

[15] Michele A. Burton, Erin Brady, Robin Brewer, Callie Neylan, Jeffrey P. Bigham, and Amy Hurst. 2012. Crowdsourcing subjective fashion advice using VizWiz: Challenges and opportunities. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility*. 135–142.

[16] CBS News. 2015. Google Apologize for Mis-Tagging Photos of African Americans. Retrieved October 8, 2015 from **http://www.cbsnews.com/news/google-photos-labeledpics-of-african-americans-as-gorillas/**.

[17] Wendy Chisholm, Gregg Vanderheiden, and Ian Jacobs. 2001. Web content accessibility guidelines 1.0. *Interactions* 8, 4 (2001), 35–54.

[18] Juliet Corbin and Anselm Strauss. 2014. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. Sage Publications.

[19] Timothy C. Craven. 2006. Some features of "alt" texts associated with images in web pages. *Information Research: An International Electronic Journal* 11, 2 (2006), 2.

[20] Maitraye Das, Darren Gergle, and Anne Marie Piper. 2019. "It doesn't win you friends": Understanding accessibility in collaborative writing for people with vision impairments. *Proceedings of the ACM on Human-Computer Interactio.* 3, CSCW (2019), Article 191, 26 pages.

[21] 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 *Proceedings of the European Conference on Computer Vision*. 15–29.

[22] 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 *Proceedings of the World Wide Web Conference*. 549–559.

[23] Cole Gleason, Amy Pavel, Xingyu Liu, Patrick Carrington, Lydia B. Chilton, and Jeffrey P. Bigham. 2019. Making memes accessible. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 367–376.

[24] Morten Goodwin, Deniz Susar, Annika Nietzio, Mikael Snaprud, and Christian S. Jensen. 2011. Global web accessibility analysis of national government portals and ministry web sites. *Journal of Information Technology & Politics* 8, 1 (2011), 41–67.

[25] 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*. 1–11.

[26] Gabriella M. Johnson and Shaun K. Kane. 2020. Game changer: Accessible audio and tactile guidance for board and card games. In *Proceedings of the 17th International Web for All Conference*. 1–12.

[27] Daniel Keysers, Marius Renn, and Thomas M. Breuel. 2007. Improving accessibility of HTML documents by generating image-tags in a proxy. In *Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility*. 249–250.

[28] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2013. Baby Talk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 12 (2013), 2891–2903.

[29] Polina Kuznetsova, Vicente Ordonez, Alexander Berg, Tamara Berg, and Yejin Choi. 2012. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 359–368.

[30] Scott C. LaBarre. 2007. ABA resolution and report on website accessibility. *Mental and Physical Disability Law Reporter* 31, 4 (2007), 504–507.

[31] Walter S. Lasecki, Phyo Thiha, Yu Zhong, Erin Brady, and Jeffrey P. Bigham. 2013. Answering visual questions with conversational crowd assistants. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–8.

[32] Eleanor T. Loiacono, Nicholas C. Romano Jr., and Scott McCoy. 2009. The state of corporate website accessibility. *Communications of the ACM* 52, 9 (2009), 128–132.

[33] 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*. 5988–5999.

[34] Alexander Patrick Mathews. 2015. Captioning images using different styles. In *Proceedings of the 23rd ACM International Conference on Multimedia*. 665–668.

[35] Valerie S. Morash, Yue-Ting Siu, Joshua A. Miele, Lucia Hasty, and Steven Landau. 2015. Guiding novice web workers in making image descriptions using templates. *ACM Transactions on Accessible Computing* 7, 4 (2015), 1–21.

[36] 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*. 1–11.

[37] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. 2010. A comparison of information seeking using search engines and social networks. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*.

[38] 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*. 5506–5516.

[39] Julius T. Nganji, Mike Brayshaw, and Brian Tompsett. 2013. Describing and assessing image descriptions for visually impaired web users with IDAT. In *Proceedings of the 3rd International Conference on Intelligent Human Computer Interaction (IHCI'11)*. 27–37.

[40] Abiodun Olalere and Jonathan Lazar. 2011. Accessibility of US federal government home pages: Section 508 compliance and site accessibility statements. *Government Information Quarterly* 28, 3 (2011), 303–309.
[41] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2Text: Describing images using 1 million captioned photographs. *Advances in Neural Information Processing Systems* 24 (2011), 1143–1151.
[42] Helen Petrie, Chandra Harrison, and Sundeep Dev. 2005. Describing images on the web: A survey of current practice and prospects for the future. In *Proceedings of the 3rd International Conference on Universal Access in Human Computer Interaction (HCII'05)*.
[43] Christopher Power, André Freire, Helen Petrie, and David Swallow. 2012. Guidelines are only half of the story: Accessibility problems encountered by blind users on the web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 433–442.
[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*.
[45] John M. Slatin. 2001. The art of ALT: Toward a more accessible web. *Computers and Composition* 18, 1 (2001), 73–81.
[46] Abigale J. Stangl, Esha Kothari, Suyog D. Jain, Tom Yeh, Kristen Grauman, and Danna Gurari. 2018. BrowseWithMe: An online clothes shopping assistant for people with visual impairments. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. 107–118.
[47] Lisa Tang and Jim A. Carter. 2011. Communicating image content. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 495–499.
[48] Anja Thieme, Cecily Morrison, Nicolas Villar, Martin Grayson, and Siân Lindley. 2017. Enabling collaboration in learning computer programming inclusive of children with vision impairments. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. 739–752.
[49] 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*. 49–56.
[50] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.
[51] Timo Volkmer, John R. Smith, and Apostol Natsev. 2005. A web-based system for collaborative annotation of large image and video collections: An evaluation and user study. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*. 892–901.
[52] 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 and Social Computing*. 1584–1595.
[53] Michele A. Williams, Caroline Galbraith, Shaun K. Kane, and Amy Hurst. 2014. "Just let the cane hit it": How the blind and sighted see navigation differently. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers and Accessibility*. 217–224.
[54] Fredrik Winberg and John Bowers. 2004. Assembling the senses: Towards the design of cooperative interfaces for visually impaired users. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*. 332–341.
[55] 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*. 1180–1192.
[56] Chien Wen Yuan, Benjamin V. Hanrahan, Sooyeon Lee, Mary Beth Rosson, and John M. Carroll. 2017. I didn't know that you knew I knew: Collaborative shopping practices between people with visual impairment and people with vision. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), Article 118, 18 pages.
[57] 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. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), Article 121, 22 pages.
[58] 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*. 2353–2362.
[59] Kathryn Zyskowski, Meredith Ringel Morris, Jeffrey P. Bigham, Mary L. Gray, and Shaun K. Kane. 2015. Accessible crowdwork? Understanding the value in and challenge of microtask employment for people with disabilities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1682–1693.

Received July 2019; revised October 2021; accepted October 2021