Going Beyond One-Size-Fits-All Image Descriptions to Satisfy the Information Wants of People Who are Blind or Have Low Vision

ABIGALE STANGL, University of Washington
NITIN VERMA, University of Texas at Austin
KENNETH R. FLEISCHMANN, University of Texas at Austin
MEREDITH RINGEL MORRIS, Microsoft Research
DANNA GURARI, University of Texas at Austin

Image descriptions are how people who are blind or have low vision (BLV) access information depicted within images. To our knowledge, no prior work has examined how a description for an image should be designed for different *scenarios* in which users encounter images. Scenarios consist of the information goal the person has when seeking information from or about an image, paired with the source where the image is found. To address this gap, we interviewed 28 people who are BLV to learn how the scenario impacts what image content (information) should go into an image description. We offer our findings as a foundation for considering how to design next-generation image description technologies that can both (A) support a departure from one-size-fits-all image descriptions to context-aware descriptions, and (B) reveal what content to include in minimum viable descriptions for a large range of scenarios.

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

Additional Key Words and Phrases: image description, image caption, alt text, blind, low vision, visual impairment, scenarios, minimum viable description, context aware

ACM Reference Format:
Abigale Stangl, Nitin Verma, Kenneth R. Fleischmann, Meredith Ringel Morris, and Danna Gurari. 2021. Going Beyond One-Size-Fits-All Image Descriptions to Satisfy the Information Wants of People Who are Blind or Have Low Vision. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '21), October 18–22, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3441852.3471233

1 INTRODUCTION

Many people who are blind or have low vision (BLV)[1] access information depicted in digital images through natural language descriptions of the images. They use such descriptions, for example, to stay up to date with the news, enjoy entertainment media, engage in social interactions, and avoid risk when sharing images they took (by learning if the image content may be deemed unprofessional, inappropriate, or private) [6, 11, 15, 50, 77, 86].

[1]Throughout this paper we use both *people first language* (people who are BLV) and *identity first language* (BLV people) depending on the grammar of the sentence, and in recognition that some people want their visual impairment acknowledged as an essential identifier and others do not.

Manuscript submitted to ACM

A challenge is how to author image descriptions[2] so that they are useful. Most efforts focus on how to produce one-size-fits-all descriptions for digital images [17, 25, 43, 49, 65, 71]. While this is a worthy goal, different people want different things in different situations, and as a result, it is challenging to find a single description that works for every user and context. Thus, a recent trend is to author image descriptions based on *context* in order to enhance their usefulness. Examples of context-aware solutions include varying the linguistic composition of an image description based on a given author's personality (e.g., dramatic, sweet, optimistic) [66], author's writing style (i.e., personal voice used on social media posts) [18], and target caption style (e.g., humorous, romantic) [26, 44]. Prior work also has identified how the information or content users' wants in a description varies based on the source where an image is found [30, 56, 70], including whether in a news website, social networking site, e-commerce website, employment website, online dating website, productivity application, or e-publication. In summary, context has been used to both determine *what content to include* as well as how to linguistically convey the visual content in a sentence structure.

Our aim is to improve understanding of how to author useful image descriptions by investigating the influence of *context*, in particular *scenarios*, on the information or *content* people who are BLV want in image descriptions, e.g., the visual elements depicted in an image (e.g., objects, people, environments, activities, and attributes of these—color, texture, gender, layout, etc.)[3]. We use the term *content wants* to refer to the visual elements that users self-select (rather than a need imposed on them by an external force [5, 13, 33, 83]) through this paper. We define *scenario* as consisting of the **information goal** the person has when seeking information from or about am image, paired with the **source** where the image is found. This definition is based on: (A) the understanding that people often have goals when using digital media that pertain to fact finding, information gathering, transactions, communication, maintenance (as opposed to more exploratory searching) [3, 35]; (B) evidence that different people who are BLV have unique content wants for the same visual media found on the same source [27]; and (C) evidence that BLV people want different content described for images found on news websites versus shopping websites versus dating websites, etc. [70].

To our knowledge we are the first to look at the information goal + source (scenarios) as a composite factor impacting image description content wants, despite evidence that going beyond one-size-fits-all image descriptions is a topic of great interest to people who are BLV, as evidenced in prior literature e.g., [19, 30, 49] and in podcasts featuring BLV technology advocates[4]. Our work is also motivated by the observation that the same image can appear on different sources e.g., [30], and images are commonly shared and re-posted from source-to-source e.g., images shared from news media to social media platforms, from web-based sources into productivity documents, and from shopping websites to social media profiles. We expect this trend of image sharing across different contexts to grow with new social and technological constructs, requiring image descriptions to be aware and responsive to both where the image is encountered and the goals of the consumer.

To address the question, *What influence do different scenarios—as an important contextual factor—have on the content BLV people want in image descriptions?*, during in-person interviews with 28 BLV people we asked them to identify they image content they wanted for *five* real images (mediated through descriptions) when presented in the context of *five* plausible scenarios. This task resulted in 700 total responses to the 25 image-scenario combinations (28 participants x 25 responses). We analyzed these responses through an exploratory process that involved qualitative inductive and deductive coding methods. In addition to exploring scenarios as a novel contextual factor, to our knowledge we are the

[2]Throughout this paper we use the term *image descriptions* interchangeably for *image captions*. For reference, "image description" is commonly used within the access community (e.g., [37]), whereas "image caption" is commonly used by computer vision scientists for the descriptions automatically created by algorithms (e.g., [25].)

[3]We exclude consideration of the linguistic challenges of how to convey the visual content in a sentence as it is out of scope.

[4]https://twimlai.com/accessibility-and-computer-vision/

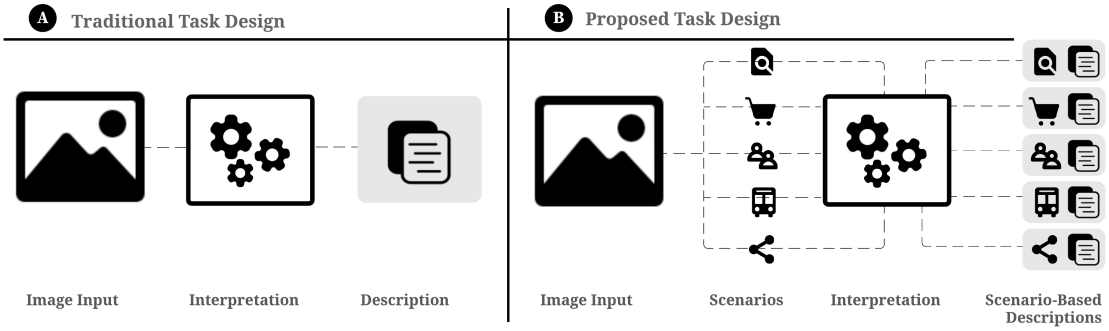

Fig. 1. Shown are two diagrams that represent both the prior and our proposed methodologies for identifying individuals' content wants in image descriptions. (A) Prior task designs took as input an image for interpretation by a human and output a description. (B) Our new task design takes as input an image and then passes it through different scenarios during interpretation to identify resulting descriptions for the same image. Application of our novel task design supports identifying (A) what content is commonly desired across a variety of scenarios and (B) what content is commonly desired only for specific scenarios.

first to investigate how context influences what content BLV people want in image descriptions in this way; prior work on the topic asked their participants to discuss their image description content wants without reference to specific images [19, 70] and/or scenarios [27, 28, 82].

In summary, our work introduces a user-centered methodology for identifying how to tailor image descriptions to specific scenarios. We conducted studies with BLV participants who followed this methodology and report our findings of the desired content for five scenarios. We conclude by discussing possible improvements to our methodology, including scaling it up for more scenarios. Our work offers a valuable foundation for improving upon the status quo for image description technologies since users reportedly can be frustrated about receiving too much information or not receiving their information of interest [27, 29, 42, 49, 58, 70].

2 RELATED WORK

2.1 Image Description Technologies

Many technologies have been introduced to deliver image descriptions. Some are human-powered [8], engaging crowdworkers [7, 58, 61, 87], friends [10], or social microvolunteers [9] to provide the descriptions. Others are fully-automated approaches that employ computer vision algorithms to describe, for example, the objects, people, and scenery in images [21, 22, 30, 38, 41, 71, 73, 75, 82]. Understanding that current automated techniques do not consistently produce accurate descriptions [42, 58], others have developed hybrid techniques that employ computer vision algorithms with humans to co-author descriptions [29, 58, 59]. While a range of services employ a one-size-fits-all approach, some researchers have questioned the usefulness of this approach for people who are BLV [49, 58, 59, 70, 71, 78], in great part because people who are BLV experience frustration when description services provide insufficient detail, too much detail, do not address their visual questions, nor help them understand the purpose of the image in context. In this paper, we offer complementary findings to both perspectives. We uncover what content (object, people, scene identification and attribute details) real users' want for a given image based on different contexts. This approach offers a valuable foundation to support the development of next-generation technologies that meet BLV peoples' preferences, both by

highlighting what content should be provided only for specific scenarios (i.e., in context-aware descriptions), and what content should be provided across all scenarios (i.e., in one-size-fits-all descriptions).

2.2 Image Description Authorship Guidelines

Prior work offers guidance in how to author image descriptions [17, 25, 34, 43, 51, 65, 67, 84]. Our work complements the subset that focuses on describing images for people who are BLV. This includes efforts by practitioners and scholars to develop the "Web Content Accessibility Guidelines" which include guidance on authorship of alternative text (alt text) [79], tools for assessing the role of the image in relationship to where it is found [16, 88], as well as templates, sample cues, and games that aid people in the description process [20, 47, 76]. Much of the prior work on this topic has relied on sighted individuals to decide how to describe images for people who are BLV [32, 82], though more and more scholars are taking human-centered approach by asking people who are BLV to share their image description preferences [19, 27–29, 70, 71]. For example, Bennett et al. (2021) ask participants how the want their and other peoples' appearance to be described and contribute important and critical perspectives on gender, race, and disability identification through description. In kind, we also take a human-centered approach, though our contribution is to reveal new image description content wants. We also demonstrate that image description guidelines can be improved to account for contextual factors, such as scenarios, which influence what content people who are BLV want in image descriptions.

2.3 BLV People's Content Wants for Image Descriptions

Prior research has offered insight into what information BLV people want in image descriptions. For example, prior work has recommended including observed objects, people, activities, location information, colors, and emotions [56] as well as the foreground, background, and directional orientation of objects [16, 68]. Much of this work has centered on social networking sites (SNS) as the context of study [27, 50, 58, 70, 77, 81, 86], and identifies that the content which should be included in descriptions for images on Facebook and Twitter includes: the name of who is in the image, where it was taken, and others' responses to it [77]; salient visual elements, people, and photo quality [86]; the number of people with each person's facial expression, pose, and age [81]; whether the image shows an inside, outdoor, or nature scene, is a close-up, or a selfie [81]; and the text and other information depicted in memes [28]. Recent work looked beyond images found on SNS to understand how BLV people's content wants change based on the source where they encounter an image [70]. Unlike prior work, we investigate how different scenarios (information goal + source) impact what content people who are BLV want in an image description. Furthermore, ours is the first work to explicitly examine BLV users' content wants for the same image across different contexts.

3 METHODOLOGY: DESCRIBING AN IMAGE ACCORDING TO DIFFERENT SCENARIOS

We now describe our approach for addressing the novel problem of identifying what content individuals want described in an image based on scenarios as a contextual factor. The key decisions we made when designing it were: (A) what scenarios to include, (B) what images to use, and (C) how to identify study participants' content wants. Towards trying to reduce participant fatigue during the interviews, we limited the number of images and scenarios to five each and so a total of 25 image-scenario combinations.

3.1 Task Design

3.1.1 Scenario Definition. We chose five *scenarios* to serve as the foundation for the task. We designed the scenarios to specify where the image is found (the source) and an information goal. Our choice of scenarios was based on those considered by prior work for examining people's content wants for different sources [70], and the understanding that people encounter images when trying to achieve a task, e.g., fact finding, information gathering, transactions, communication, maintenance [3, 35]. We authored scenarios with the aim of including enough detail for them to be plausible to the participants, yet open enough to make sense with the sample images—each of which varied in content and composition (as described below).

The scenarios we chose are as follows:

- **Scenario A** (news + learn conditions)**:** You encounter a digital image when visiting a news website [source] to learn about the working conditions as described in an article [information goal];
- **Scenario B** (e-commerce + purchase gift)**:** You encounter a digital image as you are browsing on an e-commerce website [source] to purchase a gift for yourself or a friend [information goal];
- **Scenario C** (SNS + find friend's info.)**:** You encounter a digital image when browsing your social networking site [source] with the aim to find out about one of your acquaintance's interests or activities [information goal];
- **Scenario D** (travel + plan trip)**:** You encounter a digital image when you are on a travel website [source] to help you plan a trip and learn about possible activities at the destination [information goal];
- **Scenario E** (photo library + share w/ friends)**:** You encounter a digital image that you have taken [source], that you wish to share on social media with your friends so they know what you have been doing [information goal].

3.1.2 Image Selection. Next, we selected five *sample images* to evaluate in the context of the five aforementioned scenarios. The five sample images we included in the study are shown in **Figure 2**, and we found the images from the following sources: **Image 1**, from a news media site [62]; **Image 2**, from a home shopping site [39]; **Image 3**, from a Google Images search for "outdoor work"; **Image 4**, from a travel blogger's site [57]; and **Image 5**, from a blog about a local restaurant [40]. We selected images that varied in composition as well as in subjects or objects depicted. For instance, Image 2 (Living Room) versus Image 5 (Food) had varying numbers of objects; Images 1 (Politician) and Image 4 (Mountains) showed more of the surrounding environment than other images; and the four images depicting people (Images 1, 3, 4, and 5) depicted different numbers of people engaged in diverse activities. While a challenge we faced was ensuring that each image was universally relevant across all five scenarios, our results will show that it was rare that our study participants were unable to imagine that a described image would pertain to all five scenarios.

3.1.3 Image Descriptions. We wrote a *sample image description* for each image with the understanding that each description had to serve as a surrogate through which the study participants could access information about the content of the image. Each image description was structured as a list[5] that first identified all of the people, objects, activities/interactions, and environmental features captured in the image, and then provide details about each content type. The image descriptions we provided are shown in **Figure 2**. This approach to providing descriptions while evaluating different workflows for alt-text generation is consistent with [49, 59], and Shneiderman et al.'s image visualization mantra—overview first, zoom and filter, then details [64].

[5]We chose not to present participants with a predefined sentence structure to avoid consideration of the sentence structure used to convey the content.

Image 1: Politician

People: Eight people. **Objects:** Lamp post, sidewalk, trees, cars, bicycles, bollards. **Activity/Interaction:** People walking down a sidewalk as a group; Man walking in the center, everybody else behind and to the sides of him). **Environment:** Outdoors, park, sidewalk. **People Details:** (Male 1) White skin, white hair-balding (Bernie Sanders); (Male 2) White skin with dark hair. (Male 3) White skin with light brown hair. (Woman 1) Light brown skin. (Woman 2) Light brown skin and dark long hair. (Woman 3) White skin and light brown long hair. **Object Details:** (Male 1) Black suit, white shirt, and blue tie. (Male 2) Grey suits. (Male 3) Grey suit and blue shirt. (Woman 1) Turban, polka dot shirt, black pants, and a black shoulder strap. (Woman 2) White dress with black flowers embossed on it, and a black cardigan. (Woman 3) Dress with a blue and pink floral pattern on it, and blue cardigan. **Activity/Interaction Details:** One person walks in the front and in the center of the group; all of the people are looking forward. **Environment Details:** Sidewalk approximately 12 feet wide, surrounded by grass. Background lined with large green trees. Copper blue lamppost adjacent to sidewalk. Bicycles and cars in the background.

Image 2: Living Room

People: N/A. **Objects:** Sofa, pillows, rug, coffee table, stools, fire place, basket, painting, ladder, blanket, shelf, vase, plates. **Activity/Interaction:** N/A. **Environment:** A living room arrangement decorated with furniture. **People Details:** N/A. **Object Details:** (Sofa) Blocky, light blue linen. (Pillows) Five total. Different black and white geometric patterns. (Rug) Approx. 12 X 15 feet, white with some dark lines. (Coffee Table) Light wood. triangular, three legs. (Stools) Elevated pillow, with secure base and top pad. Beige fabric. (Fire place) Modern, wood trim, black metal interior. (Basket) Hanging basket on wall with two large loop handles. (Painting) Abstract painting with white background, and swatches of colors overlayed with one another. (Ladder) Wide Navajo style ladder with blanket draped over. (Vases) Set of black, narrow neck vases. (Plates) Set of black plates with geometric pattern. **Activity/Interaction Details:** N/A. **Environment Details:** Fire place extending from back wall, creating an alcove.

Image 3: Bazaar

People: One person. **Objects:** Stall, t-shirts. **Activity/Interaction:** Man standing in front of t-shirt stall. **Environment:** Outdoors, bazaar. **People Details:** Man has dark skin, wearing orange t-shirt and brown athletic pants. **Object Details:** (Stall) Corrugated tin roof, dirt floor, rock platform. A row of 6+ t-shirts hanging, and four high stacks of different colored t-shirts that are protected by a base of blue tarps. (T-Shirts) Each shirt is a different color and pattern. **Activity/Interaction Details:** N/A. **Environment Details:** N/A.

Image 4: Mountains

People: Two people. **Objects:** Mountain, valley, clouds, river, trees. **Activity/Interaction:** Two people standing side by side on the side of a hill, overlooking a river valley. **Environment:** Valley covered with trees and clouds. **People Details:** Man and woman, wearing black winter clothing. **Object Details:** (Trees) Different colors. (River) Bends. (Mountains) Rolling. **Activity/Interaction Details:** N/A. **Environment Details:** Autumn.

Image 5: Food

People: One person's hand. **Objects:** Hand, food, plate, table. **Activity/Interaction:** Person reaching onto plate, grabbing food with right hand. **Environment:** Inside restaurant, food/plate on table. **People Details:** Man with dark skin. **Object Details:** (Food) Green vegetables, lentils, salad, injera bread. (Table) Wood. (Plate) Metal. **Activity/Interaction Details:** N/A. **Environment Details:** N/A.

Fig. 2. Images used in our interviews to examine how different scenarios impact what content people who are BLV want in image descriptions. We also show the sample image descriptions we provided to the study participants.

3.1.4 Interview Design. We designed our study to present each participant with an image description and prompt them to answer *"What content would you like in a description for the image based on the scenario we presented you with?"* for each of the five scenarios. This process would then be repeated for all five images. Consequently, each study participant would be taken through 25 image-scenario combinations. We ordered the presentation of images based on the length of the sample descriptions (starting with the image with the longest description) so that at the end of the session the study participant would have to make sense of the least amount of new information (Image 1 $\rightarrow$5). Prior to engaging them in this task we asked participants 15 open-ended questions about their visual impairment, accessible technology preferences, experience with digital images, and experience with technologies and services that provide image descriptions. A full version of our interview protocol can be found in the Supplementary Materials.

3.1.5 Data Collection. We planned to transcribe audio recordings of each interview and use each response for each image-scenario combination for subsequent analysis. Consequently, our interview protocol would yield 25 responses per participant for the 25 image-scenario combinations.

3.2 Data Analysis Approach

Given the novelty of our task design, at the outset we were uncertain what analysis methodology would be best-suited to analyze the study participants responses for all image-scenario combinations. Consequently, we explored both an inductive approach and a deductive approach to support a more comprehensive analysis from distinct perspectives.

3.2.1 Inductive Topic-Based Analysis. During data collection, we observed that the participants responded to the scenario-based task in slightly different ways. For instance, in some cases participants explained why some content was pertinent to include in the description and why other content was not important, while others repeated the list of content types and then filter out what was most important, and others simply listed what content was most important to include. Yet, a commonality we observed across most participants was that they would specify the topic they perceived to be pertinent for each image-scenario combination. For example, in response to Image 4, Scenario A a participant stated, *"It's about surroundings, what's around them. If the sidewalk is a large sidewalk or a small narrow sidewalk."* In this case we observed that the content they perceived to be the most important were the *attributes of place*. In another example for this scenario, we heard a participant state *"Take away almost everything about the humans, except for maybe their genders. 'Three guys and three women are walking down sidewalk', focus more on the scenery and what they are doing."* For this example we observed that both the *attributes of place* and *activity of people* were most important.

In turn, the lead researcher conducted an inductive thematic analysis [12, 53] of all responses to identify the range of *information topics* the participant group perceived to be pertinent for each image-scenario combination. The researcher assigned a code to each response for all 25 image-scenario combinations; each code represented a unique topic. For two of these codes *identification of scene content* and *attributes of specific content* they created-sub codes according to the types of scene, object, and person identifiers and details the participants considered to be most pertinent. Once this process was complete, they met with two other co-authors to review the code-book and examples of text corresponding to each code.[6] In some cases, the team found that a code applied for one image-scenario combination was similar to that of another image-scenario combination, but with slightly different wording, i.e. *attributes of place* versus *attributes of setting*. In situations such as these, the team reviewed the response under both image-scenario combinations, and determined that the same code could be used, i.e. *attributes of place*. In total, this inductive analysis resulted in seven main codes that were applied across the scenarios. We report both the final seven main codes, and the 28 sub-codes in **Table 1**.

3.2.2 Deductive Term-Based Analysis. We next analyzed all responses with a focus on the explicit terms participants used to specify the content they wanted. This offered a complementary approach for us to investigate participants' content wants based on the observation that in many cases the participants listed off the explicit terms they wanted to be included in the description, i.e. *"More like, trees, street, a park with large green trees, and the sidewalk"*.

Two researchers contributed to the analysis via an interactive, iterative process. Initially, both researchers manually parsed and the responses for the specific terms used by participants when indicating their content wants for Image 1 for all five scenarios. They categorized the terms according to a set of parent-codes *People, Environment, Activity,* and *Objects*, which were theoretically-derived from prior work described in Section 3.1.3. The researchers also developed inductively-derived codes by mapping the results of the inductive topic-based analysis onto the theoretical framing used for the term-based analysis.

Then, one researcher coded the participants' responses for all remaining images using the tool MAXQDA [45]. After coding was completed for each image, the second researcher reviewed the work and met with the first researcher to verify and, when needed, edit the codes. Edits typically arose because study participants could use different words to express the same/similar meaning. For instance, some participants used terminology that was more general than the terms they were presented in the sample description. Take Image 1, for example; we provided participants with a

[6]We choose not to conduct inter-rater reliability across the entire dataset because our primary goal was yield codes for each image-scenario combination, not to find agreement [46]; we found it more productive to discuss the codes identified for each image-scenario and together review the examples.

Inductive Topic-Based Analysis Main Codes and Sub-Codes			
Ab.*	*Main Codes*	*Sub-Codes*	*Definition*
ID	Identification of Scene Content...	... of people, of clothing, of furniture, of styling, of food, of landscape (e.g., mountain valley), of place (e.g., park)	Content wants center on the naming of the primary content types that make up an scene; e.g., "a couch, a chair, and a table in a room", with no attention on the appearance of the content types.
ASC	Attributes of Specific Content (People, Objects, Place)	bike, clothing, furniture, lawn furniture, plate, shirt, food styling, food, park, mountain, living room, etc.	Content wants focus on naming of content AND details that make that content distinguishable from content of the same kind; e.g., height, weight, color, pattern, style name, style detailing, cleanliness, layout [of space], arrangement [of content in space], ambiance/scenery [of space, e.g., comfort, beauty], climate.
GD	Geographic Details	Location name	Content wants center on the formal name of a given location.
ACT	Activity	People's	Content wants center on the depicted people's actions or interactions.
		Poster's	Content wants center on the depicted actions or interactions of person posting the image.
REL	Relationship	People's	Content wants center on the depicted connection between two people or a person and an object/environment depicted in an image.
		Poster's	Content wants center on the depicted connection between the person posting the image and the content in the image.
EXP	Experience	People's	Content wants center on the depicted impact of an interaction or activity has on the person depicted in the image.
		Poster's	Content wants center on the depicted impact of an interaction or activity depicted in the image has on the person posting the image (even if they are not depicted in the image).
INT	Intent	People's	Content wants center on the reason or motivation of the person dipicted in the image.
		Poster's	Content wants center on the reason or motivation of the person posting the image.
Deductive Term-Based Analysis Theoretically-Derived Codes and Inductively-Derived Codes			
	Theoretically-Derived Codes	*Inductively-Derived Codes*	*Definition*
PEO	People	Identification	(General Terms) used to identify that a person is present in the image, e.g. he, she, they, the people; (Specific Terms) used, e.g. the name of the person.
		Attributes	(General Terms) used to identify a unique aesthetic value depicted in the image, e.g. "the ethnicity of the person"; (Specific Terms) used, e.g. "an African American woman".
ENV	Environment	Identification	(General Terms) used to indicate a type of a place or setting, e.g. a park; (Specific Terms) used, e.g. "the Adirondack Mountains".
		Attributes	(General Terms) used to identify the features of a place depicted in the image, e.g. "information about the setting" ; (Specific Terms) used, e.g. "the rolling mountains".
		Location Type	(General Terms) used to specify the genre of a place depicted in the image, e.g. rural or urban.
ACT	Activity	Happening	(General Terms) used to specify that something was occurring in the image, e.g. "What is going on"; (Specific Terms) used, e.g. walking, standing, selling, hugging.
OBJ	Objects	Identification	(General Terms) used to identify that an object was present in the image, e.g. "what they are wearing"; (Specific Terms) used, e.g. "the shirts, the pants, and the tie he is wearing".
		Attributes	(General Terms) used to identify the attributes of an object, e.g. "it's color"; (Specific Terms) used, e.g. "a white shirt with black polka dots".

Table 1. List of all codes used in the "Inductive Topic-Based Analysis" and the "Deductive Term-Based Analysis".

sample description indicating the presence of a person and a white shirt with black polka dots. In cases where they used the explicit terms, e.g. polka dots, we created invivo codes using the same "polka dots" terminology. In the cases we heard participants say that they wanted to know the "what the people are wearing" we applied code the general code "what wearing". The final parent-codes and sub-codes are reported in **Table 1**; examples of the *general terms* and *specific terms* that participants used are presented under the *Definition* column.

4 STUDY IMPLEMENTATION AND FINDINGS

We now present our study implementation and findings from our analysis. Our inductive analyses resulted in a range of information topics pertinent for each image-scenario combination, and demonstrated that image descriptions need

to be context-aware based on scenarios. Our deductive analyses demonstrated that it is feasible to identify content to include in minimal viable descriptions.

4.1 Study Participants

4.1.1 Recruitment. We recruited participants by circulating an IRB-approved announcement on social media, on a listserv managed by organizations serving people who are BLV, and through snowball sampling at an independence training center. Our inclusion criteria were that participants had to be at least 18 years old, identify as BLV, and use a screen reader and/or magnification tool. The announcement explained that participants would be compensated with an Amazon gift card (20 USD per hour). At the onset we selected all respondents that met our inclusion criteria, and then after the first 20 interviews, we used purposive sampling [36] to ensure a diverse representation of gender and visual impairment.

4.1.2 Demographics. We engaged 28 people who are BLV in the scenario-based task. Altogether, the participants were demographically diverse in terms of gender (16 women, 12 men), age (18 to 67 with a mean of 39.05), education (no high school diploma to having a doctorate), and occupation (e.g., unemployed, retired, DJ, lawyer, educator). Twenty-three of the participants are US born and the rest are from India, Chinese, Ethiopia, or Armenia. All reside in the US and speak English fluently. These participants have a range of visual impairments, such as blindness from birth resulting from myopia and blindness acquired due to laser surgery. A detailed description of the 28 BLV participants is provided in the Supplementary Materials.

4.2 Data Collection

All interviews were completed in-person by one sighted researcher who audio and video recorded the conversations. All participants completed the task of identifying their content wants for all 25 image-scenario combinations in approximately one hour. Transcriptions of the audio recordings were obtained from a confidential, professional transcription service. Our resulting dataset for analysis consisted of 700 responses; i.e., 25 image–scenario combinations x 28 study participants.

4.3 Inductive Topic-Based Analysis

We report in **Table 2** the range of information topics the participant group perceived to be pertinent for each image-scenario combination. These results help to support two key findings. First, they reveal that the information topics participants wanted image descriptions to address can be similar, and, perhaps more interestingly, different across multiple scenarios for a given image, thus establishing *how image captions can be informed by scenarios*. Second, they reveal how the topics can be different across multiple images for a given scenario, thus establishing *how scenarios can transcend images*. We expand upon each finding below.

4.3.1 How Image Captions Can Be Informed by Scenarios. Our first analysis of **Table 2** is based on observing the differences in the information topics we identified, through the inductive topic-based analysis, for a given image as influenced by the different scenarios. To support this analysis, we observe each of the five rows in **Table 2** independently to examine how each of the five images were interpreted across the five different scenarios.

There are considerable differences in the themes that emerge for the same image across different scenarios. One example of this is the distinct themes that emerged for Image 1 (Politician). For **Scenario A** (news + learn conditions), participants were particularly interested in three main themes: (A) *'activity of people'*, as exemplified by a participant

	Scenario A *news+learn*	**Scenario B** *e-commerce+purchase*	**Scenario C** *SNS+find info*	**Scenario D** *travel+plan*	**Scenario E** *library+share*
Image 1 *Politician*	• attributes of park (as work setting) • attributes of people • activity of people	• attributes of clothing • attributes of bike	• intent of poster • activity of poster • experience of people • relationship of people	• identification of people • geographic detail • attributes of park • identification of park w. activity of people • identification of people w. attributes of park	• identification of park • activity of people • identification of park w. activity of people • relationship of poster w. activity of people
Image 2 *Living Room*	• identification of furniture • attributes of furniture • identification of furniture w. attributes of living room (as work setting)	• attributes of styling	• attributes of styling • attributes of living room w. intent of poster	• identification of styling • attributes of furniture • attributes of living room • intent of poster	• identification of styling • identification of living room w. identification of furniture • attributes of furniture • experience of poster
Image 3 *Bazaar*	• attributes of market (as work setting) • activity of people	• attributes of shirts	• attributes of shirt • attributes of people • attributes of market • intent of poster • intent of people	• identification of market w. activity of people • geographic detail • attributes of shirt • intent of poster	• identification of market w. activity of people • attributes of shirt • attributes of market • experience of poster
Image 4 *Mountains*	• attributes of landscape (as work setting) • activity of people	• attributes of landscape • attributes of clothing	• identification of landscape w. activity of people • geographic details • attributes of place • relationship of poster	• geographic details • attributes of landscape	• geographic details • activity of people • experience of poster • experience of people
Image 5 *Food*	• attributes of restaurant (as work setting) • attributes of food • activity of people • experience of people	• attributes of food • attributes of plate • intent of poster	• attributes of food • experience of poster	• attributes of restaurant w. identification of food • attributes of food • attributes of restaurant • experience of people (of food/of place)	• attributes of food • activity of people • experience of poster

Table 2. Shown for each image-scenario combination are all sub-themes for the types of content participants wanted.

who said *"I would probably describe like...they going to go to a meeting to talk about working conditions"*, (B) *'attributes of people'* in the image, e.g. *"I guess you do probably want to know..., like what they're wearing is probably important, also, if they're, white or African American"*, and (C) *'attributes of place [work setting]'*, for instance, *"I want to know a little bit more about the environment and... is it like, is it clean?"* In contrast, participants were interested in two distinct themes for **Scenario B** (e-commerce + purchase gift) that centered on attributes of objects in the image: (A) *'attributes of clothing'*, e.g. *"I want to know what the dresses looked like. I don't think like being in a park really matters or that Bernie Sanders is leading the pack. Let's say I wanted to buy the white dress with the black flowers, I'd want to know like, what the dress looks like and the cut of the dress stuff like that"*, and (B) the attributes of the bike, e.g. *"I would like described what color are the bikes in case somebody wanted to buy something like this"*.

Altogether, these results suggest that the image description content that people who are blind want differs based on the specific scenario in which it is encountered. In the discussion, we elaborate on how these offer a new contribution and motivate the value of designing context-aware image descriptions.

4.3.2 How Scenario Themes Can Transcend Images. Our next analysis centers on the themes that recurred under each scenario, across all of the images (columns shown in **Table 2**, e.g. Scenario A, Images 1, 2, 3, 4, and 5). Because all images depicted different content, we can observe the influence of the same scenario on the information topics identified for all five image descriptions.

Two scenarios, with very different information goals and sources, stand out as having common themes across the images. These are *visiting a news website to learn about working conditions* **(Scenario A)** and encountering an image on an *e-commerce website with the goal to purchase a gift for a friend* **(Scenario B)**. For *visiting a news website to learn about working conditions* **(Scenario A)**, participants wanted the descriptions to center on the *'activity of people'* for all of the images that depicted people (four out of five images), and *'attributes of the work setting'* for four of the five images. For encountering an image on an *e-commerce website with the goal to purchase a gift for a friend* **(Scenario B)**, participants wanted to know the *'attributes of [the primary object in the image described]'* for all five images, with the focus on clothing for Images 1 (Politician), 3 (Bazaar), and 4 (Mountains), furniture for Image 2 (Living Room), and food for Image 5 (Food). We attribute the distinct choice of what visual elements to receive attributes about to a lack of clarification about the type of the gift in the scenario; instead, participants had to self-identify the image content that could be purchased.

A slight link also exists between themes and scenario for *posting a personally-taken image on social media* (**Scenario E**). Participants shared an interest in knowing about the *'activity of poster'* for two out of five images and *'experience of poster with the activity of people'* for two out of the five images. We hypothesize that thematic differences across participants arose in part because there were differences in perceptions amongst participants regarding whether the person posting the image was also the person depicted in the image.

In contrast, there is little consistency with the themes for the scenario of *browsing to learn about an acquaintance's interests or activities on social networking platforms* **(Scenario C)** and *visiting a travel website to plan a trip* **(Scenario D)**. We suspect the thematic inconsistencies for *browsing to learn about an acquaintance's interests or activities on social networking platforms* arose because participants had to define for themselves what the activity of interest was, as opposed to relying on the scenario for this information; we authored scenarios with the aim of including enough detail for them to be plausible to the participants, yet open enough to make sense with the selected images—each of which varied in content and composition. We hypothesize if the information goal was defined to a greater degree of specificity we would have seen a stronger trend. Similarly, we suspect the thematic inconsistencies for *visiting a travel website to plan a trip* arose because the the scenario did not include information about what one would be doing on the trip. The only consistency for these scenarios what that participants wanted details about the *'intent of poster'* for two of five images when *browsing to learn about an acquaintance's interests or activities on social networking platforms*, e.g. *"That needs to be... why are they showing me this? For example, 'Because I've been in that stall"*; including these details in the scenarios may have increased consistency.

We suspect that the different outcomes observed across the different scenarios illuminate the extent to which participants' content wants were based in the images versus the scenarios themselves. We hypothesize that information goals with a higher degree of specificity would elicit greater consistency for a given scenario regarding what content participant want described about images.

4.4 Deductive Term-Based Analysis

Here we present the results from our deductive term-based analysis in two ways. First, we report how many times participants used terms related to the codes in **Table 3**. We show the frequency with which each coded term was

	Scenario A					**Scenario B**					**Scenario C**					**Scenario D**					**Scenario E**				
	news+learn					*e-comm.+purchase*					*SNS+find info.*					*travel+plan*					*library+share*				
Image	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
People																									
Identification	13	1	1	3	1	1	1	1	—	—	10	3	3	4	2	4	—	—	—	1	13	2	2	—	1
Attributes	11	1	10	4	7	9	—	2	3	1	8	—	2	6	4	1	—	7	2	3	7	—	9	3	3
Environment																									
Identification	6	1	3	3	3	—	1	1	1	1	9	—	5	4	7	4	—	11	4	5	4	2	2	5	4
Location Type	3	11	12	1	4	—	4	1	1	1	4	8	3	—	—	8	5	13	2	—	4	9	4	—	3
Milieu	3	1	—	11	5	3	1	1	9	2	4	—	3	13	—	12	2	8	18	4	11	—	2	13	3
Activity/Interaction																									
Happening	3	—	2	3	2	—	—	—	—	—	7	—	1	4	2	3	—	1	1	5	3	—	4	1	—
Specific Verbs	26	—	2	1	3	19	—	—	2	2	10	—	6	2	3	11	—	2	—	2	21	—	3	1	6
Objects																									
Identification	8	38	12	4	14	25	45	—	19	23	6	28	10	1	14	22	38	14	2	16	16	29	12	—	16
Attributes	—	10	11	—	7	43	60	52	7	23	1	21	10	3	11	—	24	13	—	15	—	11	12	1	13

Table 3. Shown is the number of times we recorded the reporting of each term across all participants' responses for each image-scenario combination.

applied across the scenarios[7]. To do this, we tallied how many study participants wanted the terms used in a description for each image-scenario combination, and then examined for each image whether the the terms were used in one, two, three, four, or five of the scenarios. We report the results in according under the parent code codes of *people, environment,* and *activity* in **Table 4**, and *objects* in **Table 5** (these results are split into two tables purely for the sake of presentation).

These results highlight the extent to which the content that participants wanted in descriptions were universal across the scenarios versus specific to a single scenario. The universal content is shown in the right-most column (i.e. 'Content Wanted in 5 Scenarios') and scenario-specific content shown in the left-most column (i.e. 'Content Wanted in 1 Scenario'). We offer this table as a reference for developers who must decide what content to include in authored descriptions, whether in general or for specific scenarios. For instance, the universal terms may be used to determine the description content to include in a *minimum viable description*, e.g. descriptions that at least contain the content that consumers need for the image to be meaningful. For images that contain people, this includes the following description content: name/identity, perceived gender, clothing style, perceived race/diversity, and facial expressions. In cases when the environment is part of the relevant information, descriptions should specify the name of a place/location, the location type (e.g. park, office, etc.), the area (how large of a space), the climate, the condition (cleanliness/safety), and information about the scenery or setting more generally. For images showing clothing for sale, color, pattern, style, material, quantity, and arrangement should be included in descriptions. Finally, for images that contain food, the participants reported wanting the taste/flavor to be described. Technology developers may use these findings to prioritize which image content to use as part of their description authoring guidelines.

5 DISCUSSION

To our knowledge, our work is the first to examine how scenarios—as important and novel contextual factor—influence the type of content people want to learn about a *single* image (as presented in **Section 2**). Below we discuss further the

[7]We did not observe trends in the participants' use of terms across images or scenarios.

Least Frequent	**Frequency of Image Description Content Wants**			*Most Frequent*
Wants for 1 Scenario	*Wants for 2 Scenarios*	*Wants for 3 Scenarios*	*Wants for 4 Scenarios*	*Wants for 5 Scenarios*
People				
(IM. 1) Height, body posture, hair color; (IM. 2) Profession; (IM. 3) Profession, Race/Diversity; (IM. 4) (none); (IM. 5) Clothing style.	(IM. 1) Physical build, age, quantity; (IM. 2) (none); (IM. 3) Age, facial expression; (IM. 4) Identity or names, Configuration of people; (IM. 5) Configuration of people	(IM. 1) Configuration of people; (IM. 2) (none); (IM. 3) Gender; (IM. 4) Facial expressions; (IM. 5) General appearance.	(IM. 1) Gender, diversity; (IM. 2) Identity or names; (IM. 3) Identity or names, general appearance; (IM. 4) Quantity, gender; (IM. 5) Identity or names, facial expression.	(IM. 1) Identity or names, Clothing style; (IM. 2) (none); (IM. 3) (none); (IM. 4) Clothing style; (IM. 5) Gender.
Environment				
(IM. 1) Setting, Landmarks, Building style, Sky color, Safety; (IM. 2) Amenities; (IM. 3) Look & Feel, Safety; (IM. 4) Landmarks, Amenities, Area; (IM. 5) Condition.	(IM. 1) Setting, Scenery, Urban or rural, Spatial configuration, Condition, Capacity; (IM. 2) Location [Room] type (Office), Tourist Spot, Setting, Capacity; (IM. 3) Setting, Parking Lot, Street, Capacity, Condition; (IM. 4) (none); (IM. 5) (none).	(IM. 1) (none); (IM. 2) Name of Place/ Location; (IM. 3) Setting, Climate, Shop/Stall; (IM. 4) Tourist Spot; (IM. 5) Restaurant, Scenery.	(IM. 1) Setting, Location type (park), Name of Place/ Location; (IM. 2) Area, Condition, Décor; (IM. 3) Location type (foreign country); (IM. 4) (none); (IM. 5) (none).	(IM. 1) (none); (IM. 2) Location type, Look & Feel; (IM. 3) Name of Place/ Location, Location type (market); (IM. 4) Name of Place/ Location, Scenery, Climate; (IM. 5) Name of Place/ Location, Look & Feel.
Activity				
(IM. 1) Bullying, Going, Leading, Lobbying, Playing music, Posing, Promoting, Outreach, Reviewing, Running, Working; (IM. 2) (none); (IM. 3) Standing; (IM. 4) Standing, Hiking, Kayaking, Picking Fruit, Walking; (IM. 5) (none).	(IM. 1) Approaching, Biking, Exercising, Gathering, Meeting, Speaking/Talking, Standing; (IM. 2) (none); (IM. 3) (nonentity); (IM. 4) (none); (IM. 5) What's Happening? (general) (general)	(IM. 1) Attending, wearing; (IM. 2) Working conditions; (IM. 3) (none); (IM. 4) (none); (IM. 5) Serving.	(IM. 1) (none); (IM. 2) (none); (IM. 3) What's Happening? (general), Selling; (IM. 4) What's Happening? (general); (IM. 5) (none).	(IM. 1) What's Happening? (general), Walking; (IM. 2) (none); (IM. 3) (none); (IM. 4) (none); (IM. 5) Eating.

Table 4. Shown in this table are the terms that were used for each of the parent codes (people, environment, activity), organized according to which image the terms were associated with, and whether the terms were frequently wanted across all five scenarios (Column E), for four scenarios (Column D), for three scenarios (Column C), for two scenarios (Column B), and for one scenario (Column A). The terms in Column E indicate the terms that may be included in a minimum viable description for the sample images, underscoring the extent to which the desired visual information is static versus scenario-specific.

contributions of our work, and provide reflections about our study design and how it may be iterated on to advance the development of next-generation image descriptions.

5.1 Towards Context-Aware Image Descriptions

We introduce a user-centered methodology for identifying how to tailor image descriptions to specific scenarios, and includes identification of desired content for five scenarios (**Table 2**). This is important since users of image description technologies reportedly can be frustrated about receiving too much information and not receiving their information of interest [27, 29, 42, 49, 58, 70]. In fact, our inductive topic-based analysis (**Section 4.3**), evidence that people who are BLV want image descriptions that are responsive to where they encounter an image (source) and the information goal

Least Frequent	**Frequency of Image Description Content Wants**			*Most Frequent*
Wants for 1 Scenario	*Wants for 2 Scenarios*	*Wants for 3 Scenarios*	*Wants for 4 Scenarios*	*Wants for 5 Scenarios*
Objects				
(IM. 1) Clothing (general), Dress, Tie, Suit, Shirt, Cardigan, Slacks, Helmet, / Park objects (general), Barn, Chair, Ducks, Trail, Office, Capital building/ Brand, Match/Fit, Patterns, Price, Size, Style, Texture, Thickness, Length, Style details, Color; (IM. 2) Brand, Condition, Weight Furniture (general), Alcove, Bed, Mantelpiece, Personal Items; (IM. 3) Stand/Kiosk, Tent, Furniture (general), Food (general), Spices, Condition, Size, Care/Maintenance; (IM. 4) Shoes, Pants, Camera / Price, Material, Brand, Size, Care/Maintenance; (IM. 5) Clothing (general), Sauces, Vegetables (general), Chicken, Coconut, Pineapples, Tomatoes, Steak, Accompaniment, Freshness, Temperature, Look/Presentation, Preparation, Smell, Condition, Vegetarian, Finger Food.	(IM. 1) Turban, Bicycles, Cars, Street, Trees; (IM. 2) Price, Care/Maintenance, Comfortableness, Portability, Quality; (IM. 3) Art, Quantity; (IM. 4) Personal Items, Jacket, Backpack, Colors; (IM. 5) Utensils, Dish, Price, Size, Colors, Quality .	(IM. 1) Lamp post; (IM. 2) Match/Fit, Size, Shape/Form, Coffee Table, Shelves, Stool; (IM. 3) Quality, Place of Manufacture, Style, Material, Brand; (IM. 4) (none); (IM. 5) Furniture (general), Restaurant type-Ethiopian, Bread/Injera, Quantity.	(IM. 1) Grass, Sidewalk; (IM. 2) Colors, Patterns, Quantity, Chairs, Ladder, Sectional, Sofa/Couch, Tables, Linen/Pillows, Utensils, Rugs/Carpets; (IM. 3) Clothing (general), Shirts; (IM. 4) (none); (IM. 5) Lentils, Salad/Greens, Taste/Flavor .	(IM. 1) (none); (IM. 2) Furniture (general), Arrangement, Sofa/Couch, Fireplace, Style, Material; (IM. 3) Colors, Patterns, Price; (IM. 4) Clothing (general); (IM. 5) Food (general).

Table 5. Shown in this table are the terms that were used for the parent-code: objects, presented in the same way as in Table 4.

they have at that time. For instance, for Image 1 (Politician) participants wanted more details (*attributes of the people or objects in images*) for Scenarios A (news+learn) and B (e-commerce+purchase), whereas they wanted the *people, objects, and scene elements* to be identified but with less detail for Scenarios D (travel+plan) and E (library+share). For Scenario C (SNS+find info.) they were focused on the *perspectives of the image poster and the people* depicted in the image.

The finding that BLV people's content wants for an image are influenced by the scenario is further evidenced by the fact our analysis revealed content types that were not previously identified when looking at source as a contextual factor alone [70]. For instance, we observe the additional influence of information goal when participants indicated that they want the descriptions to specify the connection between the people in an image and the object/environment depicted in an image, and information about the intent of the image contributor to be included in the description itself.

Moreover, we observed that the attributes that our participants wanted in an image description were specific to the information goal, e.g. learning about the attributes of a place make that make it conducive for working (spatial layout, lighting, cleanliness, safety, etc.) Scenario A.

These findings are important at a time when practitioners and scholars are trying to improve and scale human-powered, AI-powered, and hybrid description services and technologies to ensure that all images are paired with accurate descriptions, and so that people who are BLV are able to trust the image descriptions they encounter [27, 29, 42, 49, 58, 70]. Our findings may be used to improve upon existing guidelines for what information to include in image descriptions, and the creation of user-centered taxonomies used to create datasets that are *"good, high fidelity, and valid"*, e.g. how well the data explains things related to the phenomena captured by the data [60], and result in the development of automated, context-aware descriptions. We hypothesise that human and AI authored context-aware image descriptions will enable BLV people to gain access to the information they need more readily, and have the potential to mitigate BLV image consumers' frustration over the fact that existing description services provide the inaccurate, insufficient, or not situationally relevant details [49, 58, 59, 70, 71, 78].

5.2 Minimum Viable Descriptions

Both our inductive topic-based analysis (Section 4.3) and deductive term-based analysis (Section 4.4) revealed findings that could be beneficial for creating minimum viable image descriptions, thus improving techniques for authoring one-size-fits-all image descriptions (that are commonly used in practice today). For example, from the inductive topic-based analysis we observed that there were topics that spanned across images for a given scenario, despite the differences in the image content and compositions. While this was not applicable for all scenarios, we believe that this finding is promising; below we report on how one may improve the design of scenarios to observe this more broadly. From the deductive term-based analysis, we observed that there participants wanted specific types of content for all five scenarios. Accordingly, the minimum viable approach focuses on devising insight from real users what content applies across many scenarios, and so are good candidates for may different situations.

Though a critique of this approach is that that people who are BLV may want as much information about an image as possible to be able to determine the image purpose and gain equal access to the content, prior work indicates that the level of detail wanted is personal and may be influenced by a range of factors [70]. All the while, some of the information that was consistently requested across scenarios instigates ethical questions, including whether and how to convey the perceived gender and perceived race/ethnicity of people. Even when sighted viewers of an image can directly visually access and gauge such information, their judgments and assessments regarding these categories may be inaccurate. Findlater et al. (2020) suggest that additional research is needs to be conducted to assess whether potentially inaccurate information about a persons' appearance should be provided to BLV consumers of an image [23]. Addressing this call, Bennet et al. (2021) interviewed screen reader users who were also Black, Indigenous, People of Color, Non-binary, and/or Transgender on their current image description practices and preferences, and experiences negotiating theirs and others' appearances non-visually. Their findings indicate that: (1) people who are BLV engage in a variety of strategies for understanding appearance and understand that there is an important distinction between *appearance* and *identity* that needs to be considered when authoring image descriptions, and (2) that there are specific situations (e.g. scenarios) when appearance descriptions are particularly relevant [19]. These findings affirm that there can be a high emotional cost of misgendering [63] (which is one of the reasons why Facebook doesn't include this information in their automatically-generated image descriptions [82]). Methods for conveying uncertainty in image descriptions (such as through explicit presentation of error metrics [42] or implicit presentation factors such as screen reader voice or

volume) may be important areas for future study, in order to support nuanced conveyance of sensitive aspects of image descriptions. We believe that gathering many diverse stakeholders is important for determining how to appropriately interpret, describe, and label people, objects, environments, activities/interactions, their appearance, attributes and traits.

5.3 Methodological Reflections

When designing this study, we performed an extensive literature search on the topic of image description, but did not find any prior work that has attempted to identify how participants' content wants for a given image change based on scenarios. In turn, we devised an exploratory and innovative approach. Here we provide our reflections about the decisions we made during the task design and data collection, and how future work can build upon our approach.

5.3.1 Task Design Refinements:

Scenario Definition: During our investigation, we observed that there were a small number of cases when participants could not relate the image to the scenario (4.3% = 30/700 responses). In these instances, we heard participants ask for additional information. For example, when responding to Scenario B (e-commerce + purchase gift) for the different images, we heard participants trying to discern which objects within the image were available for purchase and who they were buying it for. This also included a participant who wanted to know the ingredients in the food (Image 5, Scenario D), *"...'cause I'm allergic to coconut and pineapples,"* evidencing a personal information goal that would help them to make sense of the scenario in relation to the image, and ultimately receive a useful image description. Based on participants' feedback, we hypothesise that scenarios can be made more accurate with more details about their information goals. To identify these details, in future work we recommend including BLV in identification of their information goals when encountering images on different sources.

Sample Image Selection: For this study we chose five images, each presenting a different composition and scene. Though our findings are useful, others who choose to draw on our approach may want to select images depicting either a wider or a narrower range of content. This could include conducting studies with different genres of images such as memes, GIFs, mixed media art, illustrations, diagrams, and/or maps.

Image Description and Response Task: Though we designed the sample descriptions using insights from prior work [49, 59, 64], we cannot dismiss the possibility that the terms used in the sample description, as well as the amount and order of the terms presented may have affected participants' responses. We made the following observations that may be used to inform future iterations of the response task.

First, we observed several notable trends regarding the participants' response to the terms we provided in the image descriptions in relation to the given scenarios. While participants often adopted the terms provided in their responses to each image-scenario combination, we observed that some people also introduced new terminology or terms that were adjacent to the terms we provided (e.g., instead of "sofa" they would say "couch") or they would use more general terms to cluster image content of the same kind (e.g., 'all of the clothing' instead of specifying specific articles of clothing. Second, we observed that in some cases participants specified terms outside of what was actually depicted in the image. For instance, from the term-based analysis, for Image 1 we heard several participants ask for a "duck" to be included in the description, though there was not a duck depicted in the image. Another description content type that fell outside of what was depicted in the image included "price," and emerged predominantly for the scenarios of browsing on an e-commerce website (Scenario B) and planning a trip (Scenario D). Though a subjective interpretation of worth may be

inferred from an image, this information is not commonly included in image descriptions. We suspect the participants relied on their prior experiences interacting with/in the people, objects, scenes, and other cultural phenomena as cognitive reference points to infer what content "might/should" be in the image. Third, we observed other instances when participants wanted more subjective or a personal assessment of the environment, object, or person, as opposed to simply an overview of the visual elements depicted in the image. For instance, for Image 5, participants wanted to know the taste of the food. In other cases, participants wanted image descriptions to include the motivations or backgrounds of the people of the image, motivations of the description author (i.e., photographer and/or poster of the image).

We also observed a trend regarding the participants' response to the amount of terms in the sample descriptions. Namely, in some cases we heard participants state that there was too much detail, that the provided descriptions were repetitive (e.g. naming all objects and then going through and naming the attributes), while in other cases they asked additional questions about the information provided in the image description. We did not hear participants indicate that the order in which we presented the content within the sample description was problematic. However, to assess for bias future work may randomize the order in which the content is presented to each participant.

Based on these observations, we make several hypothesizes and recommendations that can be used to improve upon our task design. First, with respect to the authorship of the sample descriptions, an alternative approach would be to simply name or identify the primary content depicted in the image, and prompt participants to ask visual questions about the details of the content that they perceived as important (in the context of their information goal, the source, and the surrounding media). This approach may create more consistency in how participants respond to the image description and task. We make this suggestion based on participants response to the terms (as described above). We also share several hypothesise that may be investigated to advance this work (A) There are differences in how individuals make sense of the visual content presented in the same image description; (B) A persons' prior experiences and cognitive reference points may impact their sense-making of the image descriptions or visual questioning. We leave an investigation into BLV peoples' sense-making of auditory or textual descriptions of visual content to future work, as well as investigation into how people who are BLV want phenomenological and sensory experiences like taste and color to be described, as well as how to account for the personality of the image consumer and the image author in descriptions.

5.3.2 Data Collection.

Randomization: With respect to the presentation of the images and the scenarios to participants, we chose an order meant to minimize participant fatigue, by presenting the most verbose image descriptions first to all participants; we determined that randomization and counterbalancing were less critical since this study was not designed as a controlled experiment. Still, several responses indicated fatigue, such as when we heard participants say *"what you said before"* and *"I want a lot of detail"* in some of their responses. We saw that there were a larger number of such responses for Image 5 (Food). In future studies, the order of images and scenarios could be randomized, especially if the lengths of the sample descriptions could be kept relatively consistent, as variation in length or complexity of sample descriptions could be a confounding factor.

Multimodality: At the onset of the study we opted not to provide Braille or tactile graphics of the images along with the audio descriptions because they are typically not available in real-time as people encounter digital images. That is because of a scarcity of readily-available and affordable tactile graphics and interfaces for screen-based devices

[69]. While tactile experiences are known to support image comprehension [14, 24, 54, 80], none of our participants asked for these materials during the interviews. Future work may explore the benefits of using multimodal materials as mediators.

5.3.3 Data Analysis.

Inductive Topic-Based Analysis. When presenting our findings, we focused on each theme as equal in weight. However, for some image-scenario combinations there were many more themes and in some cases there was great distribution in the numbers of participants whose responses we coded according to those themes. While we chose to represent the diversity of themes in this study, future work may benefit from devising a strategy to determine which scenarios elicit uniform versus varied responses across participants. In doing so, subsequent work could then uncover why there was more/less unity in their responses; we hypothesize that there are multiple approaches to sense-making of auditory presentations of visual information and that an individual's experiences and reference points may be another valuable contextual variable that influences what to include in image descriptions.

5.4 Future Work

Though our findings represent an important first step to creating next-generation image descriptions, more work is needed to identify what content to include in image descriptions based on the context with larger image datasets, scenarios, and participants. This may be done using the task design refinements described above. In future work we also aim to further identify and assess the influence of other factors on how to author context-aware image descriptions, e.g., the nature of the text and media surrounding the image, the immediacy of the information want, or personal experiences (prior visual experience, media and information literacy), through a factorial vignette survey experiment [4][8]. Third, future work may focus on how to deliver context-aware image descriptions through screen recognition techniques, e.g., [85], and voice-user interfaces [2], for images and videos. Finally, though it was beyond the scope of this paper, more work is needed to evaluate how to linguistically tie the content into sentences, in what order to present the content (e.g., all at once versus iteratively from high-level overview to fine-grained detail), and what delivery mechanism to use (e.g. audio vs. Braille). To the best of our knowledge these considerations have not yet been explored, but can be guided by our research.

6 CONCLUSION

In this paper we demonstrate a new approach for identifying the content—or visual elements depicted in an image—that BLV people want to be included. Unlike prior work, we take into account the *scenario* (information goal + source) in which a person who is BLV encounters an image. The methods and findings we present may be used by scholars and practitioners who are working to refine the ways in which image descriptions are generated by human-powered services, AI-powered services, and hybrid services. We offer this work as a valuable step to improve the accessibility of images for people who are BLV, as well as more broadly anyone interested in improving the quality and responsiveness of image descriptions to context. More broadly, our findings contribute to the emerging body of research focused on providing guidance for image description authorship based on the lived experiences of people who are BLV [19, 70, 71], and calls to ensure that people with disabilities are included in the development of automated technologies [1, 19, 31, 48, 52, 55, 70, 74].

[8]Vignette experiments consists of set of systematically varied descriptions of subjects, objects, or situations in order to elicit respondents' beliefs, attitudes, perceptions, or intended behaviors with respect to the presented vignettes [72]

7 ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable input, study participants for their involvement, and Jaxsen Day for his support with data analysis. This work is supported by Microsoft gift funding.

REFERENCES

[1] We Count: Fair Treatment, Disability and Machine Learning. (????). https://www.w3.org/2020/06/machine-learning-workshop/talks/we_count_fair_treatment_disability_and_machine_learning.html

[2] Ali Abdolrahmani, Kevin M. Storer, Antony Rishin Mukkath Roy, Ravi Kuber, and Stacy M. Branham. 2020. Blind Leading the Sighted: Drawing Design Insights from Blind Users towards More Productivity-oriented Voice Interfaces. *ACM Transactions on Accessible Computing* 12, 4 (Jan 2020), 1–35. DOI: http://dx.doi.org/10.1145/3368426

[3] Anne Aula and Daniel M Russell. 2008. Complex and exploratory web search. In *Information Seeking Support Systems Workshop (ISSS 2008), Chapel Hill, NC, USA*. Citeseer.

[4] Katrin Auspurg and Thomas Hinz. 2014. *Factorial survey experiments*. Vol. 175. Sage Publications.

[5] Wendy Beautyman and Andrew K Shenton. 2009. When does an academic information need stimulate a school-inspired information want? *Journal of Librarianship and Information Science* 41, 2 (2009), 67–80.

[6] Cynthia L Bennett, Martez E Mott, Edward Cutrell, Meredith Ringel Morris, and others. 2018. How Teens with Visual Impairments Take, Edit, and Share Photos on Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 76.

[7] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and others. 2010. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 333–342.

[8] Jeffrey P Bigham, Richard E Ladner, and Yevgen Borodin. 2011. The design of human-powered access technology. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*. ACM, 3–10.

[9] Erin Brady, Meredith Ringel Morris, and Jeffrey P Bigham. 2015. Gauging receptiveness to social microvolunteering. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1055–1064.

[10] Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P. Bigham. 2013. Visual Challenges in the Everyday Lives of Blind People. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 2117–2126.

[11] Stacy M Branham, Ali Abdolrahmani, William Easley, Morgan Scheuerman, Erick Ronquillo, and Amy Hurst. 2017. Is Someone There? Do They Have a Gun: How Visual Information about Others Can Improve Personal Safety Management for Blind Individuals. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 260–269.

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

[13] Anita Brown. 2004. Reference services for children: information needs and wants in the public library. *The Australian Library Journal* 53, 3 (2004), 261–274. DOI: http://dx.doi.org/10.1080/00049670.2004.10721654

[14] Emeline Brule, Gilles Bailly, Anke Brock, Frédéric Valentin, Grégoire Denis, and Christophe Jouffrais. 2016. MapSense: multi-sensory interactive maps for children living with visual impairments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 445–457.

[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*. ACM, 135–142.

[16] Diagram Center. 2015. Specific Guidelines: Art, Photos & Cartoons. http://diagramcenter.org/specific-guidelines-final-draft.html. (September 2015). (Accessed on 02/12/20).

[17] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017a. Attend to You: Personalized Image Captioning With Context Sequence Memory Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[18] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017b. Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 895–903.

[19] Morgan Klaus Scheuerman Jeffrey P. Bigham Anhong Guo Cynthia L. Bennett, Cole Gleason and Alexandra To. 2021. "It's Complicated": Negotiating Accessibility and (Mis)Representation in ImageDescriptions of Race, Gender, and Disability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

[20] Dhruba Dahal. 2018. *Simplifying and Improving Effectiveness of Image Description for Accessibility using Sample Cues*. Master's thesis. OsloMet-Oslo Metropolitan University.

[21] Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. *arXiv preprint arXiv:1505.01809* (2015).

[22] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, and others. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1473–1482.

[23] Leah Findlater, Steven Goodman, Yuhang Zhao, Shiri Azenkot, and Margot Hanley. 2020. Fairness issues in AI systems that augment sensory abilities. *ACM SIGACCESS Accessibility and Computing* 125 (2020), 1–1.

[24] Giovanni Fusco and Valerie S Morash. 2015. The tactile graphics helper: providing audio clarification for tactile graphics using machine vision. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*. 97–106.

[25] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017a. StyleNet: Generating Attractive Visual Captions With Styles. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[26] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017b. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3137–3146.

[27] Cole Gleason, Amy Pavel, Himalini Gururaj, Kris Kitani, and Jeffrey P. Bigham. Making GIFs Accessible. https://www.colegleason.com/static/papers/MakingGIFsAccessible_ASSETS2020.pdf. (Accessed on 09/08/2020).

[28] Cole Gleason, Amy Pavel, Xingyu Liu, Patrick Carrington, Lydia B. Chilton, and Jeffrey P. Bigham. 2019. Making Memes Accessible. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*. Association for Computing Machinery, Pittsburgh, PA, USA, 367–376. DOI: http://dx.doi.org/10/ggt2d6

[29] Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M. Kitani, and Jeffrey P. Bigham. 2020. Twitter A11y: A Browser Extension to Make Twitter Images Accessible. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. DOI: http://dx.doi.org/10.1145/3313831.3376728

[30] 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*. ACM, 518.

[31] Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, and Meredith Ringel Morris. 2019. Toward Fairness in AI for People with Disabilities: A Research Roadmap. *arXiv preprint arXiv:1907.02227* (2019).

[32] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *European Conference on Computer Vision*. Springer, 417–434.

[33] Philip Hider. 2006. Search goal revision in models of information retrieval. *Journal of information science* 32, 4 (2006), 352–361.

[34] Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, and others. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1233–1239.

[35] Melanie Kellar, Carolyn Watters, and Michael Shepherd. 2006. A goal-based classification of web information tasks. *Proceedings of the American Society for Information Science and Technology* 43, 1 (2006), 1–22.

[36] Paul J. Lavrakas. 2008. Purposive Sample, Encyclopedia of Survey Research Methods. Sage Publications, Inc., 2008.. (2008).

[37] Veronica Lewis. 2018. How to Write Alt Text and Image Descriptions for the visually impaired. (Jan. 2018). https://www.perkinselearning.org/technology/blog/how-write-alt-text-and-image-descriptions-visually-impaired Accessed on: 2020-05-05.

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

[39] Livingspaces.com. 2020. Living Room Ideas & Decor. https://www.livingspaces.com/inspiration/rooms/living-room-ideas. (2020). (Accessed on 02/16/2020).

[40] Mopay Lola. 2019. Final Idea. https://medium.com/@mopaylola/final-idea-5090ad49bb2a. (May 2019). (Accessed on 02/16/2020).

[41] Christina Low, Emma McCamey, Cole Gleason, Patrick Carrington, Jeffrey P Bigham, and Amy Pavel. 2019. Twitter A11y: A Browser Extension to Describe Images. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 551–553.

[42] 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*. ACM, 5988–5999.

[43] Alexander Mathews, Lexing Xie, and Xuming He. 2015. SentiCap: Generating Image Descriptions with Sentiments. (2015).

[44] Alexander Patrick Mathews, Lexing Xie, and Xuming He. 2016. Senticap: Generating image descriptions with sentiments. In *Thirtieth AAAI conference on artificial intelligence*.

[45] MAXQDA. 2020. All-In-One Tool for Qualitative Data Analysis & Mixed Methods. https://www.maxqda.com/. (2020). (Accessed on 02/18/2020).

[46] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[47] 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 (TACCESS)* 7, 4 (2015), 12.

[48] Meredith Ringel Morris. 2020. AI and accessibility. *Commun. ACM* 63, 6 (May 2020), 35–37. DOI: http://dx.doi.org/10.1145/3356727

[49] 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*. ACM, 59.

[50] 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*. ACM, 5506–5516.

[51] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. *arXiv preprint arXiv:1603.06059* (2016).

[52] Karen Nakamura. 2019. My Algorithms Have Determined You're Not Human: AI-ML, Reverse Turing-Tests, and the Disability Experience. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*. Association for Computing Machinery, 1–2. DOI: http://dx.doi.org/10.1145/3308561.3353812

[53] Kimberly A Neuendorf. 2019. 18 Content analysis and thematic analysis. *Advanced Research Methods for Applied Psychology* (2019), 211.

[54] Valeria Occelli, Charles Spence, and Massimiliano Zampini. 2008. Audiotactile temporal order judgments in sighted and blind individuals. *Neuropsychologia* 46, 11 (2008), 2845–2850.

[55] Joon Sung Park, Danielle Bragg, Ece Kamar, and Meredith Ringel Morris. 2021. Designing an Online Infrastructure for Collecting AI Data From People With Disabilities. (2021), 12.

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

[57] David Clapp Photography. 2020. Portfolio Categories. https://www.davidclapp.co.uk/portfolio. (2020). (Accessed on 02/16/2020).

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

[59] Elliot Salisbury, Ece Kamar, and Meredith Ringel Morris. 2018. Evaluating and Complementing Vision-to-Language Technology for People who are Blind with Conversational Crowdsourcing.. In *IJCAI*. 5349–5353.

[60] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. (2021), 15.

[61] Christine Samson, Casey Fiesler, and Shaun K. Kane. 2016. "Holy Starches Batman!! We Are Getting Walloped!": Crowdsourcing Comic Book Transcriptions. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '16)*. Association for Computing Machinery, New York, NY, USA, 289–290. DOI: http://dx.doi.org/10.1145/2982142.2982211

[62] Bernie Sanders. 2019. This Is How We Will Cancel All Student Debt. https://medium.com/@SenSanders/this-is-why-we-should-cancel-all-student-debt-6ea987d02ce2. (June 2019). (Accessed on 02/16/2020).

[63] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article Article 144 (Nov. 2019), 33 pages. DOI: http://dx.doi.org/10.1145/3359246

[64] Ben Shneiderman. 1996. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*. 336–343. DOI: http://dx.doi.org/10/fwdq26 ISSN: 1049-2615.

[65] Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Engaging Image Chat: Modeling Personality in Grounded Dialogue. *CoRR* abs/1811.00945 (2018). http://arxiv.org/abs/1811.00945

[66] Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019. Engaging image captioning via personality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 12516–12526.

[67] Rachel N Simons, Danna Gurari, and Kenneth R Fleischmann. 2020. " I Hope This Is Helpful" Understanding Crowdworkers' Challenges and Motivations for an Image Description Task. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.

[68] John M Slatin and Sharron Rush. 2002. *Maximum accessibility: Making your web site more usable for everyone.* Addison-Wesley Longman Publishing Co., Inc.

[69] Abigale Stangl, Ann Cunningham, Lou Ann Blake, and Tom Yeh. 2019. Defining Problems of Practices to Advance Inclusive Tactile Media Consumption and Production. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 329–341. DOI: http://dx.doi.org/10.1145/3308561.3353778

[70] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. DOI: http://dx.doi.org/10.1145/3313831.3376404

[71] 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*. ACM, 107–118.

[72] Peter M. Steiner, Christiane Atzmüller, and Dan Su. 2017. Designing Valid and Reliable Vignette Experiments for Survey Research: A Case Study on the Fair Gender Income Gap. *Journal of Methods and Measurement in the Social Sciences* 7, 22 (Jun 2017). DOI: http://dx.doi.org/10.2458/v7i2.20321

[73] 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.

[74] Shari Trewin. 2018. AI Fairness for People with Disabilities: Point of View. *arXiv:1811.10670 [cs]* (Nov 2018). http://arxiv.org/abs/1811.10670 arXiv: 1811.10670.

[75] 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.

[76] 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 (CHI '04)*. Association for Computing Machinery, New York, NY, USA, 319–326. DOI: http://dx.doi.org/10.1145/985692.985733

[77] Violeta Voykinska, Shiri Azenkot, Shaomei Wu, and Gilly Leshed. 2016. How blind people interact with visual content on social networking services. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 1584–1595.

[78] Alexandra Vtyurina and Adam Fourney. 2018. Exploring the role of conversational cues in guided task support with virtual assistants. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 208.

[79] W3C Web Accessibility Initiative. 2018. Web Content Accessibility Guidelines (WCAG) Overview. (May 2018). https://www.w3.org/WAI/standards-guidelines/wcag/ Accessed on 2020-05-05.

[80] Zheshen Wang, Baoxin Li, Terri Hedgpeth, and Teresa Haven. 2009. Instant tactile-audio map: enabling access to digital maps for people with visual impairment. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*. 43–50.

[81] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2016. Ask Me Anything: Free-Form Visual Question Answering Based on Knowledge from External Sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4622–4630.

[82] 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*. ACM, 1180–1192.

[83] Bo Xie. 2009. Older adults' health information wants in the internet age: Implications for patient–provider relationships. *Journal of health communication* 14, 6 (2009), 510–524.

[84] Xiaoyu Zeng, Yanan Wang, Tai-Yin Chiu, Nilavra Bhattacharya, and Danna Gurari. 2020. Vision skills needed to answer visual questions. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–31.

[85] Xiaoyi Zhang, Lilian de Greef, Amanda Swearngin, Samuel White, Kyle Murray, Lisa Yu, Qi Shan, Jeffrey Nichols, Jason Wu, Chris Fleizach, and others. 2021. Screen Recognition: Creating Accessibility Metadata for Mobile Applications from Pixels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[86] 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), 121.

[87] 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*. ACM, 2353–2362.

[88] Ying Zhong, Masaki Matsubara, and Atsuyuki Morishima. 2018. Identification of Important Images for Understanding Web Pages. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 3568–3574.