UCLA
UCLA Electronic Theses and Dissertations

Title

Netizen A11y: Engaging Internet Users in Making Visual Media Accessible

Permalink

https://escholarship.org/uc/item/6gg864hq

Author

Wang, Ruolin

Publication Date

2024

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Netizen A11y: Engaging Internet Users in Making Visual Media Accessible

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Electrical and Computer Engineering

by

Ruolin Wang

2024

© Copyright by
Ruolin Wang
2024

ABSTRACT OF THE DISSERTATION

Netizen A11y: Engaging Internet Users in Making Visual Media Accessible

by

Ruolin Wang
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2024
Professor Xiang 'Anthony' Chen, Chair

Visual media without alternative text descriptions are inaccessible to blind and low vision (BLV) people. While mainstream automated solutions often focus on achieving acceptable results based solely on visual media, they neglect the essential engagement of internet users in contributing to accessibility. With an explosion of data, diverse media formats, and constant updates, the responsibility of ensuring accessibility and equity cannot rest solely on the shoulders of a select few. The sheer volume of digital contents demands a community effort where each individual becomes an advocate for accessibility.

With the central goal of effecting this community-level change, I design and develop intelligent user interfaces that engage internet users in making visual media accessible at three progressive levels: i) extracting visual information from human-authored texts in metadata to generate in-context descriptions; ii) nudging people to use visual descriptive language in online activities; iii) creating symbiotic information seeking experiences that connect sighted and BLV people of the same community.

Three practical application scenarios, online shopping, online video watching and commenting, online research paper reading, are explored to demonstrate the implementation

of each engagement level. Moving beyond paid crowdsourcing, this dissertation envisions a future where user advocacy plays a crucial role in creating a more unified and inclusive internet.

The dissertation of Ruolin Wang is approved.

Jon E. Froehlich

Kai-Wei Chang

Nanyun Peng

Gregory J. Pottie

Xiang 'Anthony' Chen, Committee Chair

University of California, Los Angeles

2024

To my mom

I couldn't have done this without you.

TABLE OF CONTENTS

1 Introduction **1**

1.1 The Vision of Netizen A11y: Every Internet User Can Become an Advocate for Accessibility 1

1.2 Opportunities and Challenges of the Three-level Engagement 2

1.3 Document Outline and Summary of Contributions 4

1.4 Authorship Statement 7

2 Extracting Visual Information from Human-Authored Texts: An Application for Online Shopping **8**

2.1 Introduction 8

2.2 Related Work 10

2.2.1 Online Shopping Accessibility 11

2.2.2 Information Retrieval In Online Reviews 12

2.3 Overview of Studies 14

2.4 Understanding Current Practice and Deriving Design Implications 15

2.4.1 Challenges 15

2.4.2 Current Solutions 16

2.4.3 Implications for Design 17

2.5 Leveraging Reviews to Answer BLV Customers' Questions 19

2.5.1 Product Categories and the Corresponding Question Types 19

2.5.2 Types of Visual Questions and How to Find Informative Answers . . 20

2.5.3 Rules-Based Solution for Retrieving Informative Review Snippets . . 22

2.5.4 Generating Answers and Image Descriptions 24
2.6 Implementation of Revamp—An Information Retrieval System for Online Shopping 25
2.7 Evaluation 28
2.7.1 Rule Evaluation 29
2.7.2 System Evaluation 31
2.8 discussion 35
2.8.1 Working with Multiple Product Pages of Online Shopping 36
2.8.2 Working with A Broader Variety of Product Pages 36
2.8.3 Working with Supervised methods 36
2.8.4 Working with Better QA Experience 37
2.8.5 Working in the Real-World Settings 38
2.8.6 Limitations 38
2.9 Conclusion 39

3 Nudging People to Use Visual Descriptive Language: An Application for Online Video Watching and Commenting 40
3.1 Introduction 40
3.2 Related Work 43
3.2.1 Utilizing Human-Authored Texts to Enhance Accessibility 43
3.2.2 Authoring Tools for Writing Descriptions 45
3.2.3 Nudging for Social Good 46
3.3 The Design of Nudges at Video Watching and Commenting Time 47
3.3.1 Visualizations of Inaccessible Video Segments 48

3.3.2 Hints on Describing Inaccessible Segments 49
3.3.3 Prompts of Making Comments Accessible 50
3.3.4 Captions and Comments as References 50
3.4 Implementation of CoSight: A Chrome Extension On YouTube 51
3.4.1 Identifying and Explaining Inaccessibility 51
3.4.2 Analyzing and Storing Comments 52
3.5 Study with Sighted Viewers: Does CoSight Nudge People to Write Accessible Comments? 53
3.5.1 Methods 53
3.5.2 The Usability of CoSight Interface for Sighted Viewers 54
3.5.3 The Content of Comments Written with CoSight 56
3.6 Interviews with Experts: Do Comments Written with CoSight Really Work? 59
3.6.1 Method 60
3.6.2 Results 61
3.7 Discussion 65
3.7.1 Nudging People with Limited Knowledge with Scaffolding 65
3.7.2 Nudging People with Limited Motivation with Gamification 66
3.7.3 Potential and Concerns of Involving AI in the Loop 66
3.7.4 Values and Challenges of Deploying in the Wild 67
3.7.5 Limitation and Future Work 68

4 Creating Symbiotic Information Seeking Experience: An Application for Online Research Paper Reading 69

4.1 Introduction 69

4.2 Related Work . . . 71
4.2.1 Reading Tools for General Users . . . 71
4.2.2 Reading Tools for Blind and Low Vision Users . . . 72
4.2.3 LLM-based Techniques in Assisting Research Reading . . . 73
4.3 Method . . . 74
4.3.1 Participants . . . 74
4.3.2 Protocol and Data Analysis . . . 76
4.3.3 Design Probe . . . 76
4.3.4 Think-Aloud with the Demo . . . 78
4.4 Results . . . 79
4.4.1 Understanding Sighted Researchers . . . 79
4.4.2 Understanding BLV Researchers . . . 83
4.5 Discussion . . . 87
4.5.1 Mapping Needs of BLV and Sighted Researchers . . . 87
4.5.2 Building Collaborative Intelligence of AI and Crowd . . . 89
4.5.3 Crafting a Unified Tool for All Users with Different Abilities in One Community . . . 90
4.5.4 Limitation and Future Work . . . 91

5 Summary . . . 93
5.1 Limitations and Future Work . . . 94
5.2 Closing Remark: Towards a Unified and Inclusive Internet . . . 95

A Appendix to Chapter 2 . . . 96
A.1 Demographics of study participants . . . 96

B Appendix to Chapter 3 . . . . . **98**

B.1 Instructions on writing accessible comments with CoSight 98

B.1.1 Why your comments may help blind people 98

B.1.2 Make your comment more accessible by clarifying the visual 98

B.1.3 Write a descriptive comment, especially for an inaccessible segment . 99

B.2 Explaining inaccessible segments and Instructions on What to Describe . . . 100

B.3 Comments Collected by CoSight in Study with Sighted Audience 100

C Appendix to Chapter 4 . . . . . **102**

C.1 Example 1: Blind Researcher Interacting with chatGPT 102

C.1.1 Round 1 102

C.1.2 Round 2 103

C.1.3 Round 3 105

C.2 Example 2: Blind Researcher Interacting with chatGPT 106

References **110**

LIST OF FIGURES

2.1 Revamp simplifies and reconstructs the original Amazon web page for Blind or Low Vision users' information seeking task to understand product details, especially the appearance. Using rule-based heuristics, Revamp extracts descriptive information from customer reviews to generate image descriptions (a), responses to users' queries (b) with a sentiment summary of all the relevant reviews (c) and original reviews sorted into a positive and a negative lists (d). 10

2.2 System overview of how Revamp responds to users' query. It first extract the keywords and process the query by its kind. After filtering out irrelevant reviews and reranking, Revamp uses the shortest candidate among the top-3 reviews to generate the visual description of a product and relevant answers to the query. 25

2.3 An example questionnaire for BLV users to vote for which one of the review snippets (retrieved by our rules or by default ranking of search result) best answers visual attributes questions. 28

2.4 Performance of image description and visual questions on four basic attributes including shape, logo, size, color. Two sighted research team members rated the quality by three levels on 15 Amazon top-seller products in each of the three main categories. Results showed that our rules can directly provide visual info for 54 questions or image descriptions in Clothing, Shoes & Jewelry; 40 in Home & Kitchen; 29 in Electronics. We also included "Not Applicable" because of no relevant reviews or too few informative reviews. 30

2.5 Examples of representative products in three main categories, showing the retrieved reviews, generated image description and comparing with the recognition results by SeeingAI, a common tool for BLV users to figure out what is in an image. 33

2.6 Subjective ratings on statements comparing current practice with experience of using Revamp. Revamp demonstrates a clear advantage in the experience of understanding the product appearance. 34

3.1 CoSight, a Chrome extension that augments YouTube video pages with interactive features to nudge the sighted viewers to contribute to video accessibility by anchoring and grounding their comments to specific visual contents of the videos. Blind and low vision users of CoSight can listen to these time-anchored descriptive comments of visuals as they play the videos. 42

3.2 Features applying on video page, at watching time including 1) Visualizations of Inaccessible Video Segments; and 2) Hints on Describing Inaccessible Segments; at commenting time including 3) Prompts of Making Comments Accessible; and 4) Captions and Comments as References. 48

3.3 The interaction flow of writing a descriptive comment for an inaccessible segment: hovering on accessibility icon to trigger the pop-up card with hints on describing inaccessible segments, clicking on the icon to jump to input box with timestamp automatically added, then write the descriptions with related caption and comments as references. 49

3.4 When the user write a comment to share their opinions, reminders on adding timestamps and clarifying ambiguous pronouns are shown based on the input content to nudge the user to anchor and ground their comment with specific visual detail. 50

3.5 Participants reported ratings from 1 to 7on the user experience of CoSight, including the following statements: 1) How mentally demanding was writing accessible comments? (1 = low, 7 = high) 2) How distracting of writing accessible comments to the video watching experience? (1 = low, 7 = high) 3) After using this tool, I understand the concept of video accessibility. (1 = highly agree, 7 = high disagree) 4) This tool motivates me to contribute to the accessibility of online videos. 5) This tool makes it easier for me when writing accessible comments. 6) If these features are built in Youtube, I will use this tool to make my comments more accessible when watching videos in the future. 54

3.6 The histogram of counts of comments written per participant. Each participant watched two videos. 57

3.7 Analysis on the content of 257 comments showed that participants using CoSight included timestamps and descriptive visual details in their comments. Participants mainly wrote comments to describe labeled inaccessible segments, while they also wrote comments to share opinions with clarified visual referents. . . . 58

4.1 The demo activates a sidebar in the right side, supporting figure-centric assitive functions including highlights and annotations, questions, and summary. The left side shows the original HTML paper webpage. 77

LIST OF TABLES

2.1 We collected participants' subjective ratings on Revamp. The scale was 1 to 7, in which 7 = I strongly agree with this statement, 4 = neutral, and 1 = I strongly disagree with this statement. The value represents the average of ratings (SD). Data was analyzed using Wilcoxon test and a statistical significant difference ($p < 0.05$) is marked with $*$—all ratings of Revamp significantly outperformed the current practice. 32

3.1 Demographics of blind participants. 61

4.1 Demographics of BLV and sighted participants in the study. All participants read research papers often and regularly, more than one paper every week and have research experience in their fields over 3 years. 75

A.1 The demographic information of participants. 97

B.1 The accessible comments with timestamps and descriptive visual details written by participants. 101

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to:

- my family, especially my mom, for her endless love and unshakable trust in me; and my grandmother, who is the intrinsic motivation of my research journey on visual impairments;
- my PhD advisor, Xiang 'Anthony' Chen, for his full support, enabling, empowering and inspiring me to explore and work on the most important research questions I am passionate about;
- my committee members, Gregory J. Pottie, Nanyun Peng, Kai-Wei Chang, and Jon E. Froehlich, for their valuable advice and support;
- my mentors, colleagues, collaborators and mentees who worked with me, as I am always the one who benefit a lot from these experiences and connections;
- my Master advisors, Yuanchun Shi and Chun Yu, for introducing me into the field of Human-Computer Interaction;
- my "pears" (peers): my dearest friends who always cheer me up and give me hugs in the darkest times;
- my blind and low vision friends who believe in the value of my work and bring light into my life, and all study participants who make this research possible.

VITA

2011–2015 B.S. (Electronic Science and Technology), Tianjin University

2016–2019 M.S. (Computer Science and Technology), Tsinghua University

2019–2024 Ph.D. Candidate (Electrical and Computer Engineering), University of California, Los Angeles

PUBLICATIONS

Ruolin Wang. **Netizen A11y: Engaging Internet Users in Making Visual Media Accessible**. *In Proceedings of the 2024 ACM Conference on Intelligent User Interfaces.*

Noyan Evirgen, ***Ruolin Wang***, Xiang 'Anthony' Chen. **From Text to Pixels: Enhancing User Understanding through Text-to-Image Model Explanations**. *In Proceedings of the 2024 ACM Conference on Intelligent User Interfaces.*

Xingyu "Bruce" Liu, ***Ruolin Wang***, Dingzeyu Li, Xiang 'Anthony' Chen, Amy Pavel. **CrossA11y: Identifying Video Accessibility Issues via Cross-modal Grounding**. *In Proceedings of the 2022 ACM symposium on User Interface Software and Technology.*

Xuhai Xu, Tianyuan Zou, Han Xiao, Yanzhang Li, ***Ruolin Wang***, Tianyi Yuan, Yuntao Wang, Yuanchun Shi, Jennifer Mankoff, Anind K Dey. **TypeOut: Leveraging Just-in-Time Self-Affirmation for Smartphone Overuse Reduction**. *In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.*

Ruolin Wang, Zixuan Chen, Mingrui Zhang, Zhaoheng Li, Zhixiu Liu, Zihan Dang, Chun Yu, Xiang 'Anothony' Chen. **Revamp: Enhancing Accessible Information Seeking Experience of Online Shopping for Blind and Low Vision Users**. *In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.*

Mingrui 'Ray' Zhang, ***Ruolin Wang***, Xuhai Xu, Qisheng Li, Ather Sharif, Jacob O. Wobbrock. **Voicemoji: Emoji Entry Using Voice for Visually Impaired People**. *In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.*

Zihan Wu, Chun Yu, Xuhai Xu, Tong Wei, Tianyuan Zou, ***Ruolin Wang***, Yuanchun Shi. **LightWrite: Teach Handwriting to The Visually Impaired with Only A Smartphone.**. *In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.*

Ruolin Wang, Chun Yu, Xing-Dong Yang, Weijie He, Yuanchun Shi. **EarTouch: Facilitating Smartphone Use for Visually Impaired People in Public and Mobile Scenarios**. *In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.*

CHAPTER 1

Introduction

1.1 The Vision of Netizen A11y: Every Internet User Can Become an Advocate for Accessibility

The accessibility of digital information has become a rising concern due to the prosperity of internet and social media. In our daily interactions with digital technologies, visual media is a primary, even preferred, mode of delivering and accessing information. However, for users with disabilities, images, videos, diagrams, charts, and even emojis can inhibit rather than enhance access to information. Where a set of stairs without ramps might pose a barrier for people with limited physical mobility, visual media without alternative text descriptions is inaccessible to Blind and Low Vision (BLV) people, as well as those with dyslexia, cognitive impairments, and age-related macular degeneration. So discrepancies in access that we see in the physical world are also translated into the digital world, causing digital inequities. With an explosion of data, diverse media formats, and constant updates, the responsibility of ensuring accessibility and equity cannot rest solely on the shoulders of a select few, *e.g.*, crowdworkers [1], friends [2], volunteers [3], and visual content creators [4]. The sheer volume of digital content demands a community effort where every internet user can become an advocate for accessibility.

The term *netizen*, as *net citizen*, describes an individual who actively seeks to contribute to the improvement of the internet. My dissertation points to a future of ***Netizen A11y*** [1],

[1] A11y is an abbreviation for accessibility with 11 representing the count of letters between a and y.

where social and crowd intelligence are central contributions to online information accessibility. To facilitate this community-level change, I design and develop intelligent user interfaces that engage internet users in making visual media accessible at three progressive levels:

- **Level I**: Leveraging information that people unintentionally and randomly contribute to accessibility;
- **Level II**: Nudging people to intentionally and proactively contribute to accessibility, with additional effort;
- **Level III**: Creating social information-seeking experiences where people naturally contribute to accessibility, aligned with their original goals.

Although this framework may engage and benefit people with mixed abilities, my dissertation focuses on two groups: people who are blind and low vision, and people with sight. Individuals with almost complete vision loss are considered blind or legally blind. Those who have partial sight may prefer terms such as low vision. In the following chapters, I will use the term BLV people and sighted people to refer to these two groups.

1.2 Opportunities and Challenges of the Three-level Engagement

Researchers have explored different approaches of generating visual descriptions, including fully human-authored (*e.g.*, [5]), fully automated (*e.g.*, [6]) and hybrid techniques of computer vision algorithms co-authoring with humans (*e.g.*, [7]). While most mainstream solutions focus on achieving acceptable results solely based on the visual media, they often overlook the varying needs of users depending on the image's source [8]. Human-authored texts on social platforms, although not specially written for describing the visual media, are always neglected as a useful information source for generating in-context image descriptions. For example, in online shopping websites such as Amazon, the customer reviews may contain vivid details of visual appearance of the products.

At **Level I**, to leverage information that people unintentionally and randomly contribute to accessibility, my work offers novel insights into accessible information seeking by ***extracting visual information from existing human-authored texts*** in the source where the visual media is found. Such visual information can be used to generate vivid descriptions for visual media based on first-hand experiences from real humans who directly consume and interact with the media. Sighted people are passively involved by providing the data they publish online, such as reviews and comments. BLV people are provided with an alternative to the original online platform, augmented with interactive features to support accessible information seeking.

The effectiveness of the first-level approach necessitated a substantial amount of human-authored texts to specify visual details. However, without the awareness of intentionally contributing to accessibility, only a small proportion of these human-authored texts contain informative descriptions for BLV users. For instance, based on our observations, it may be challenging to extract enough informative reviews from an unpopular online product on Amazon with fewer than one hundred reviews. Similarly, video comments on YouTube often lack context when referencing visuals, making interpretation challenging. Moreover, videos, as a form of visual media, possess a higher information density compared to static images, thus requiring an even greater volume of informative comments to bridge the visual information gap. This presents an opportunity to enhance people's engagement by nudging for accessibility. HCI Researchers have applied the concept of nudging [9] to guide users towards desired behaviors in multiple contexts for social good, including health, sustainability and privacy. Despite this, we still lack an understanding of applying effective nudges for contributing to accessibility.

At **Level II**, to nudge people to intentionally and proactively contribute to accessibility, my work explores the design of lightweight interventions by ***nudging sighted people to use visual descriptive language in online posts***. Such nudges not only actively engage more sighted individuals in enhancing visual media accessibility but also foster broader

societal awareness and involvement in accessibility efforts. Both sighted and BLV people are provided with an alternative to the original online platform, augmented with different interactive features to contribute to accessibility or benefit from these social contributions.

The effectiveness of the second-level approach hinges on people's willingness and intrinsic motivation to spend additional effort contributing to accessibility. Rather than imposing additional burdens beyond their original goals, it is worth investigating whether accessibility contributions can be seamlessly integrated into activities that individuals are already motivated to perform. Considering the needs of end-users with disabilities can lead to the creation of accessible tools that benefit everyone. However, this also presents unprecedented challenges in designing symbiotic information-seeking experiences that meet and reconcile the shared and differing goals of various user groups.

At Level III, to create social information-seeking experiences where people naturally contribute to accessibility, my work embeds the value of accessibility into social information foraging by ***designing a unified tool that benefits both sighted and BLV individuals*** within the same community. This approach ensures that accessibility becomes an integral part of the information-seeking process, also fostering collaboration and mutual support among all community members, regardless of their abilities. Both sighted and BLV individuals are provided with a unified tool that enhances their information-seeking experience.

1.3 Document Outline and Summary of Contributions

As outlined above, I explore three threads of work under three levels of sighted user engagements in this dissertation: *(i)* extracting visual information from existing human-authored texts; *(ii)* nudging sighted people to use visual descriptive language in online posts; *(iii)* designing a unified tool that benefits both sighted and BLV people. My dissertation presents three representative application scenarios, including online shopping, online video watching

and commenting, online research paper reading, each aligned with the corresponding level.

The outline of this dissertation is as follows:

- **Chapter 2** presents Revamp, an interactive information retrieval system [10] that integrates with Amazon and leverages customer reviews to support accessible information seeking of BLV users. Online shopping has become a valuable modern convenience, but BLV users still face significant challenges using it, because of: *(i)* inadequate image descriptions, and *(ii)* the inability to filter large amounts of information using screen readers. To address those challenges, Revamp supports product browsing with review-based image descriptions and question-answering interactions on a reconstructed product page. From our interviews, we identified four main aspects (color, logo, shape, and size) that are vital for BLV users to understand the visual appearance of a product. Based on the findings, we formulated syntactic rules to extract review snippets, which were used to generate image descriptions and responses to users' queries. Evaluations with eight BLV users showed that Revamp *(i)* provided useful descriptive information for understanding product appearance, and *(ii)* helped the participants locate key information efficiently.

- **Chapter 3** presents CoSight, a Chrome extension to augment YouTube video pages with interactive nudging features to support accessible comments authoring of sighted audience. The explosive growth of online videos has exacerbated the accessibility gap for BLV individuals in need of Audio Description (AD), which cannot be met timely only through the efforts of professional AD writers. Human-authored reviews and comments can be used to describe visual media, while prior research revealed the deficiency of descriptive visual details in video comments and elicited the potential to nudge sighted viewers to contribute to video accessibility by anchoring and grounding their comments to specific visual contents. To achieve this goal, we designed and developed interactive nudging features including visualization and pop-ups at watch-

ing time, and prompts and references at commenting time. Study with sighted users showed that CoSight integrates contributing to accessibility into video watching and commenting flow intuitively, and 89% of the comments written with CoSight contain descriptive visual details. Study with AD professionals and BLV people showed that these comments can provide BLV people with helpful visual information and also convey the vibes and emotions that are crucial to understand the videos.

- **Chapter 4** introduces Colla, a social reading tool that works on HTML formats of the Association for Computing Machinery (ACM) Digital Library [2]. Understanding figures in research papers is essential for education and professional research careers, yet the majority of online papers lack informative figure descriptions. Automated methods for generating scientific figure descriptions and crowdsourcing tools that seek help from random sighted individuals often fail to fully meet the expectations of BLV researchers. To address this, we propose enhancing the accessibility of paper figures through social reading, which connects and benefits both BLV and sighted researchers in similar fields. Figure-centric reading behaviors such as highlighting, annotating, and question-answering not only improve sighted researchers' comprehension but also feed into human-in-the-loop AI systems to create informative and interactive figure descriptions for blind researchers. To explore this idea, we conducted a mixed-methods study involving interviews and think-aloud sessions with sighted and BLV researchers from diverse backgrounds, using Colla as a design probe. The concept was well-received by both user groups, demonstrating its potential for further investigation and real-world deployment. We discuss the lessons learned and future directions for developing a unified tool that facilitates equitable information exchange among community members.

- **Chapter 5** concludes this dissertation, summarizing the key contributions and findings, reflecting on the limitations of the current work, and outlining future directions.

[2] A database of full-text articles and bibliographic literature covering computing and information technology.

My dissertation contributes to the human-computer interaction literature with a three-level framework for engaging internet users in making visual media accessible. This framework can be generalized to various social media types and platforms, beyond the three exemplar scenarios covered. My research holds practical value with the potential to elicit platform-level changes, which could significantly impact a large portion of real-world users who require alternative text descriptions for visual media to access online information equally. With this dissertation, I argue that researchers should not only focus on accommodating specific needs by developing assistive technologies but also thoughtfully consider how to foster a culture where accessibility and equity are embedded in the fabric of digital interactions.

1.4 Authorship Statement

Although I led the research detailed in this dissertation, it is the result of joint efforts and contributions from my advisor and mentors, my colleagues and collaborators, as well as all the sighted and BLV participants in my studies. To acknowledge my collaborators' contributions, I use the first-person plural in the following chapters. Part of my work is also built upon prior works from my colleagues on identifying accessibility issues in videos [11, 12].

CHAPTER 2

Extracting Visual Information from Human-Authored Texts: An Application for Online Shopping

2.1 Introduction

Online shopping has gained increasing popularity among Blind or Low Vision (BLV) people who have limited mobility to travel to physical stores. According to a research survey conducted in the UK, over 90% of people with disabilities shopped online at least once a month [1]. The recent COVID-19 pandemic further accelerated the adoption of online shopping among BLV users [2]. Thus making online shopping experience accessible has become an imperative requisite for ensuring the quality of life for BLV users.

However, prior studies [13, 14] show that BLV people still face significant barriers on online shopping websites due to inadequate image descriptions and screen readers' inability to filter out a large amount of information on a product page. Our formative study with 20 BLV people showed that automatic tools (*e.g.*, Seeing AI, filters of screen reader) were only used by a few experienced users, and the information provided was too generic to inform a purchase decision, especially for certain categories where the appearance matters, *e.g.*, fashion products. The most efficient way for BLV users is still seeking help from a sighted person, such as a family member or a crowdsourced helper, which is not always available.

[1] http://www.clickawaypound.com/index.html

[2] https://www.shropshirestar.com/news/health/coronavirus-covid19/2020/04/07/online-shopping-priority-plea-for-the-blind/

To address the above challenges, we developed *Revamp*, an interactive information retrieval system that supports review-based question-answering (QA) on a reconstructed product page. It is implemented as a browser extension for Amazon.com as shown in Figure 4.1. Utilizing reviews as a knowledge source, Revamp extracts informative descriptions to help BLV users understand the product appearance. To understand what kinds of visual questions BLV users care about, we collected questions from ten BLV participants on their frequently shopping categories on Amazon (Clothing, Shoes & Jewelry, Home & Kitchen, and Electronics), identified four main aspects (color, logo, shape, and size) in understanding the appearance. Based on the results, we formulated syntactic rules to extract the review snippets to generate image descriptions and responses to users' queries.

We tested the performance of these rules on the Amazon best-seller product lists covering three main shopping categories. Results showed that these rules *(i)* covered 85% informative reviews voted by BLV participants and *(ii)* provided insights on information conveyed by images evaluated by two sighted people. We evaluated the usability of Revamp with eight BLV people on six representative products in the three aforementioned categories. Participants reported that Revamp reduces their efforts of information seeking, improves their utilization of reviews, and extracts reviews that are informative to help them understand product appearance. They considered Revamp to be helpful for shopping online independently when no sighted helpers were available.

Our **contributions** are:

- Understanding the challenges faced by BLV people in information seeking when shopping online and design implications to meet their unique needs;
- Identifying specific questions important to BLV users when shopping online and deriving syntactic rules that retrieve informative reviews to provide image description or to answer visual questions;
- Revamp, an interactive information retrieval system that integrates with Amazon to

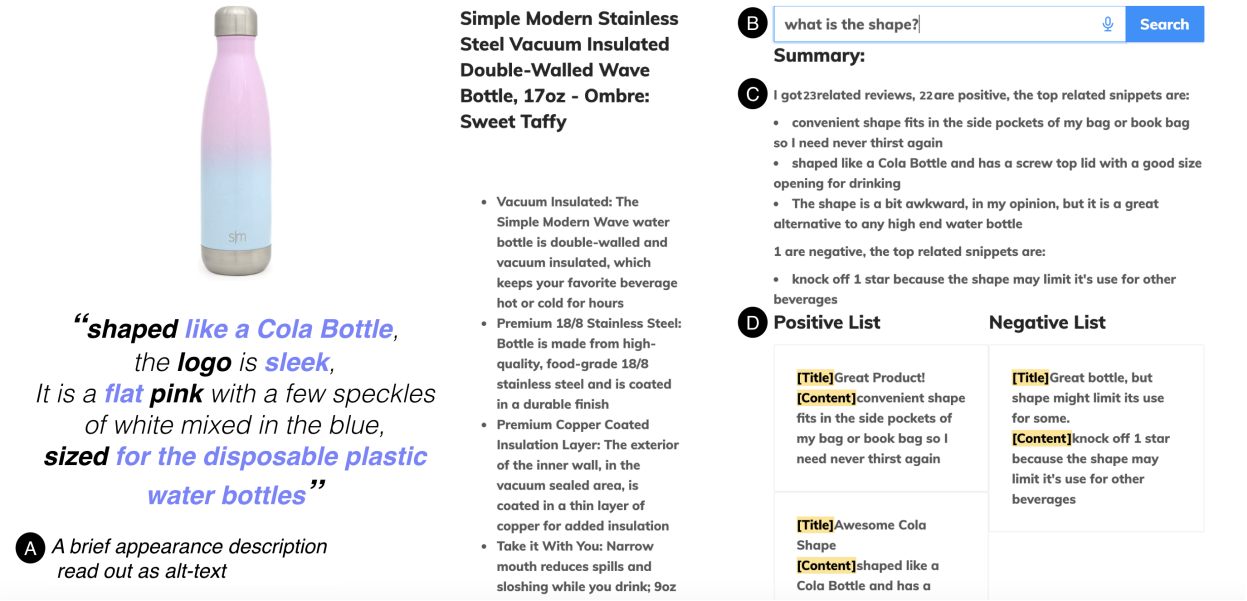


Figure 2.1: Revamp simplifies and reconstructs the original Amazon web page for Blind or Low Vision users' information seeking task to understand product details, especially the appearance. Using rule-based heuristics, Revamp extracts descriptive information from customer reviews to generate image descriptions (a), responses to users' queries (b) with a sentiment summary of all the relevant reviews (c) and original reviews sorted into a positive and a negative lists (d).

support product browsing and review-based question-answering;

- Evaluation study with eight BLV users that validated the feasibility and usefulness of enhancing accessible online shopping experience via Revamp.

2.2 Related Work

This research builds upon prior work from different sub-disciplines across accessibility, information retrieval (IR), computer vision (CV) and natural language processing (NLP). We

first review previous research investigating online shopping experience for BLV users, as well as the efforts to improve the accessibility of such experience; we then provide an overview of information retrieval methods that utilize online reviews to generate useful information, which serves as the foundation of the implementation of Revamp.

2.2.1 Online Shopping Accessibility

Blind or low vision (BLV) people often take great effort to find and learn what products are visually appropriate for them [15, 16]. Without real-time communication with store clerks and touching the products directly, information related to online products is mainly conveyed through images and text provided by sellers. Hence, inadequate descriptions of images and unparsed text information significantly reduce BLV consumers' engagement with online shopping websites [13, 14]. Currently, BLV people could leverage human-powered assistive tools to answer visual questions about daily objects and products in stores [17, 18, 2, 19, 20, 21, 22]. Specially in shopping clothes, human helpers can provide remote assistant on describing articles of clothing (*e.g.*, color) [18, 2] or offering subjective fashion advice [19, 22]. Relatedly, automatic photo caption generation and visual question answering in computer vision and AI research is often designed for general scenarios [23, 24, 25]. There exist automatic solutions that promote accessible image understanding on the web, *e.g.*, providing image alt-text for photos in Facebook based on object recognition [26], using OCR for photos on Twitter and other websites alike [27, 7], and using reverse image search to find existing captions [28].

However, such solutions cannot fulfill BLV users' special need in better understanding visual details of online products. Most images on eCommerce websites, if captioned by emerging automatic caption tools such as SeeingAI [1], will only have a generic description of the product, *e.g.*, "*probably a close up of a red chair*", without any specific insights on

[1] https://www.microsoft.com/en-us/ai/seeing-ai

appearance details, *e.g.*, how is the size, and how does this kind of red make people feel. BrowseWithMe [14] has made an important step forward in describing clothes outfit by identifying the image regions, but only respond with basic color names (*e.g.*, , "*Brick Top, Navy Pants*") based on color detection of image. Our contribution is providing a novel perspective of leveraging reviews as an external information source to supplement the visual details.

Besides inadequate descriptions of images, the online shopping experience of BLV people is also hindered by the weaknesses of screen readers dealing with crowded websites. Screen readers provide fine-grained information navigation and control, but at the cost of reduced walk-up-and-use convenience [29]. Prior work augmented the auditory web browsing experience by adding a secondary voice output [30, 31], supporting basic queries about the information provided by sellers [14], and supporting users to choose how many and which levels of detail to listen to based on their interest [32]. voice assistants as an alternative solution lack the ability to deeply engage with content and to get a quick overview of the landscape (e.g., list alternative search results & suggestions) [29]. Current voice assistants for online shopping such as *Alexa* only respond to a limited range of queries, *e.g.*, "*add bananas to the chart*", rather than finding targeted information to answer a BLV user's question. Inspired by VERSE [29] which extends a voice assistant with screen-reader-inspired capabilities to enhance web search, our work integrates review-driven question-answering with two levels of details (summary & original review lists), which contains rich first-hand insights from other buyers.

2.2.2 Information Retrieval In Online Reviews

Most image captioning and visual question answering benchmarks to date focus on questions such as simple counting and object detection that are directly based on the images. Marino *et al.* [33] draws on external knowledge resources when the image content is insufficient to answer customers' questions, but their focus on images containing general scenes cannot

be directly leveraged for online shopping. Online reviews have been mentioned as an useful resource to assist BLV customers' desire for more product information [14, 15, 13], such as supporting question-answering[34, 35] and generating summaries [36, 37]. Prior work towards general users investigated extracting experiences [38, 39], tips [40, 41] or snippets suitable for product descriptions [42, 43], yet it remains unknown how reviews can help with the unique interests of BLV users, especially providing descriptive information on product appearance. Our research fills a gap in the literature by leveraging existing human-authored resources of online reviews as an additional source of information to address such unique needs.

The most related work [42] applies a supervised method to extract reviews for image description based on 25K labeled sentences. However, the descriptions generated by this work are generic and rarely covered descriptions of visual details. Rather than following supervised methods such as conducting aspect-based sentiment analysis utilizing feature engineering [44, 45] or Bidirectional Encoder Representations from Transformers (BERT) [46] that requires a huge amount of human-labelled data, we proposed empirically-formulated syntactic rules based on our studies with BLV participants to retrieve meaningful review snippets for understanding appearances. Compared to "black box" data-driven approaches, these syntactic rules are explainable and can be validated by our targeted users; they are modular—existing rules can be edited or removed and new rules can be added without affecting the others. To the best of our knowledge, there is no prior work that directly informs rule design on extracting visual descriptions from reviews.

In summary, Revamp extends prior work on enhancing accessible information seeking by proposing a solution to bridge customer reviews with the needs of BLV users on visual information for a wider range of online products, including three frequently shopped categories: Home & Kitchen, Clothing, Shoes & Jewelry, and Electronics.

2.3 Overview of Studies

In this work, we conducted three groups of studies:

- **Formative study (Section 4, 5)** with two stages to: (i) understand the current practice and derive design implications via 30-mins semi-structured interviews with 20 participants (P1-P20); (ii) investigate what specific visual information BLV users expect and how reviews can help via questionnaires and 30-mins semi-structured interviews with 10 participants (P11-P20).
- **Rule evaluation (Section 7.1)** with two stages to: (i) evaluate the informativeness of reviews extracted by the rules via questionnaires with eight participants (P13-P20); (ii) evaluate the rules' generalizability on different products where two sighted researchers rated the quality of the generated alt-text and answers in three informative levels.
- **System evaluation (Section 7.2)** to evaluate how the Revamp prototype affected the shopping process. Eight participants (P13-P20) browsed products with Amazon web pages and Revamp then provided subjective comments by online interviews. The study lasted for 40 mins.

The time interval between each group of study was about one month. Twenty participants (Gender: 7 female and 13 male; Age: average = 27.7, SD = 5.5) came from two different cultures (P1-P10: China; P11-P20: North America) participated in the first-stage formative study due to the distributed nature of the research team. Although the platforms and languages they use differ, participants share similar screen reader experience and information seeking challenges since the screen readers follow the general design and most online shopping websites are similarly structured with overloaded information. Only North American users were involved at the following studies because of the limited availability of review data on the Chinese shopping sites. Participants were compensated with \$15 USD\hour. Each study was audio recorded, transcribed and coded by three of the authors following the reflexive

thematic analysis methods from Braun and Clarke [47]. The demographic information of participants is attached as Appendix.

2.4 Understanding Current Practice and Deriving Design Implications

We conducted 30-minute semi-structured online interviews with 20 BLV consumers (P1-P20) recruited through social media platforms. Specifically, the goal of this study is to understand that, when shopping online:

- What challenges do BLV users face in information seeking and what are their coping strategies?
- What are the design heuristics if we want to build an accessible information retrieval system for BLV users?

2.4.1 Challenges

Aligned with prior work [13, 14], two main challenges mentioned by our participants are lack of visual information and lack of efficient ways to navigate through information. We revealed how the two challenges were intertwined with each other, which elicited insights on designing an efficient way to leverage reviews to fill in the lacking visual information.

2.4.1.1 Lack of Visual Information

The most frequently mentioned challenge is the lack of detailed information for understanding the visual appearance of the product (18 out of 20 participants). While sighted users can tell the visual attributes such as color, shape and size or even functionalities of a product from a single image, the alt-texts of many current product images are either empty, or set

as the image path, which does not provide any visual information for BLV users at all. Sellers usually provide no textual descriptions equivalent to the visual information conveyed by images, whether it is some basic attribute such as color and shape, or some vivid details such as pattern design. Take the color attribute as an example: the names provided by sellers sometimes can be too generic or obscure to interpret accurately. So much of the language we use to describe a product is centered around visual, *e.g.*, "*marble pattern* " or "*Arctic blue* ", which poses a barrier to people who don't experience the world visually. P18 brought up an example that a T-shirt with the color name "*surf the web* " was actually a kind of "*bright blue* ", which was only mentioned by a review.

2.4.1.2 Lack of efficient ways to navigate through information

Comparing with the scanning and skimming experience of sighted users, the passive and sequential reading manner of screen readers makes it inefficient for BLV users to retrieve useful information scattering all over a web page. P16 mentioned that it can be frustrating when accidentally jumping into another unrelated web component such as the shopping cart in the process of exploring product details. All participants agreed that reviews were useful resources. Review snippets with detailed descriptions can also provide further insights for product appearance. However, most participants (16 out of 20) reported their utilization of reviews was greatly reduced because sifting through the large amount and unrelated reviews can be especially laborious. P19 stated that "*Usually there are too few reviews per page and there are no easy ways to jump from review to review. I know there exist useful reviews but it's hard to directly pull them out without efforts.*".

2.4.2 Current Solutions

There still exists a gap between current solutions and BLV users' expectations of online shopping experience. Relying on human helpers, the same as many other scenarios, suffer from

the limitation of availability, raises privacy concerns, and reduces BLV users' independence. Meanwhile, as reviewed earlier in Related Work, many existing techniques lack adaptation to the specific needs of BLV users' online shopping.

2.4.2.1 Seeking Help from Sighted People

80% of the participants chose to seek help from sighted people including families/friends, other customers, paid video chat and accessibility hotlines. In this way, they had to trust the helper's personal judgment and subject to issues such as helpers' inavaliability and privacy concerns. P19 mentioned that "*The shipping and delivery process is not the most time-consuming for me. Waiting for someone to help me with selecting products is.* " P12 stated that he did not always want others to know about certain products he buys and hoped there was an alternative solution to shop more independently.

2.4.2.2 Leveraging Existing Web Functions and Automatic Tools

P18 mentioned using visual interpretation apps such as Seeing AI to get brief generic descriptions on product images based on object and color detection, *e.g.*, "*probably a close up of a red chair* ". Two participants who were more tech-savvy in web interfaces (P11, P20) mentioned using keyword-based search and filters, but these tools are under-utilized by other participants and can only remove irrelevant information on a general basis. P8 and P13 mentioned the customer questions & answers section supported by websites can also help with some generic details, yet it seldom covers appearance-related information, which is assumed available to sighted consumers via product images.

2.4.3 Implications for Design

Based on the interview results, we derived 3 design implications to improve the accessibility of online shopping experience. Specifically, the designs focus on leveraging the existing resources

on a product page: *simplifying the webpage* (4.3.1) and *responding to active queries* (4.3.3) are to address the *lack of efficient ways to navigate through information* (4.1.2); *leveraging reviews to supplement visual description* (4.3.2) is to address the *lack of visual information* (4.1.1).

2.4.3.1 Simplifying and reorganizing the webpage structure

When the users' current task is understanding a product, especially the appearance, the overloaded information and web components can become burdens for seeking visual-related description. To address this problem, we can provide an alternate view of the original page with less related information and components removed, such as shopping cart. Meanwhile, it is important to provide users the option to switch back to the original page.

2.4.3.2 Leveraging customer reviews for extra visual description

Participants' responses suggest that some subjective information in reviews can be helpful for BLV people to understand the visual attributes. Just as P2 stated: "*It is impossible for me to build a visual concept as you sighted people do, but I can feel what you feel. Reviews talking about the feelings when seeing the appearance are helpful for me to understand this pattern.*" It remains to be investigated how to retrieve the informative reviews to meet the specific needs of BLV users.

2.4.3.3 Responding to active queries with the page resources

Participants expressed preferences to actively asking questions and getting summative answers of a product, which was usually what they did when consulting with a sighted helper. Meanwhile, any additional assistance mechanisms should be compatible with users' current screen reader experience to support users to access the original information as well. P6 commented "*I do not completely trust in answers selected by machines in case it may not*

cover very well ".

2.5 Leveraging Reviews to Answer BLV Customers' Questions

The first-stage formative study uncovered the opportunities and challenges to support the unique needs of BLV population in visual information seeking in online shopping. Building upon these findings, we then seek to further understand:

- What specific visual or other product-related questions are BLV users interested in?
- How can we retrieve useful review snippets to answer these questions?

It is beyond the scope of this paper to exhaustively cover the large number of product categories and all the consumer questions corresponding to each category. To narrow down our focus, we first distributed questionnaires to ten North American participants (P11 - P20) who mainly use Amazon to collect their most frequently shopped product categories, then conducted one-on-one interviews to formulate the understanding of their specific questions on three main categories: Home & Kitchen, Clothing, Shoes & Jewelry, and Electronics, based on which we derived rule-based solutions to extract informative review snippets.

2.5.1 Product Categories and the Corresponding Question Types

The frequently shopped categories on Amazon by our participants are: Electronics (chosen by 90% participants), Home & Kitchen (50%), Pet Supplies (50%), Audible Books and CDs (50%), Clothing, Shoes & Jewelry (30%), Grocery and Gourmet Food (30%). Among these categories, we focused on three main categories whose appearances are as crucial as other information for making purchase decisions, including Home & Kitchen, Clothing, Shoes & Jewelry, and Electronics.

We then gathered 168 questions from participants in these three categories. Each participant was shown representative products on the Amazon best-sellers list and was asked to

raise as many questions they want to ask as possible after browsing the title and description provided by sellers. The questions were then labeled and divided into two groups, non-visual questions and visual questions. The distinct breakout of visual (57%) and non-visual (43%) questions is based on whether a question can be answered by directly observing the image.

First, 72 out of the 168 questions were questions on non-visual information that shares common interests with sighted consumers, including *(i)* functionality and other non-visual attributes, *e.g.*, material, price; and *(ii)* consumer feedback, *e.g.*, whether the product was worth the price, or of high quality.

We are interested in the remaining 96 questions on visual information, which were of specific interests to BLV users, including *(i)* visual attributes of image, *e.g.*, color, logo, shape, size; and *(ii)* high-level concepts that can be inferred from an image, *e.g.*, usage method, style.

2.5.2 Types of Visual Questions and How to Find Informative Answers

For non-visual questions, all participants reported that the useful information usually could be found in the product description provided by the sellers, reviews or customer questions & answers, while for visual questions, they mainly rely on human helpers to get the answers. Thus we focus our investigation on the following sub-categories of visual questions and whether we can provide informative answers to these visual questions based on reviews, hence providing an alternative for BLV users to shop online independently.

2.5.2.1 Visual questions on basic attributes

Most participants started with general appearance questions, such as "*how does it look like?* ", followed by detailed questions around the basic visual attributes of the product, including color, shape, size and logo, which puts forward the need of providing a briefing of appearance covering these basic attributes first, then responding to specific queries. Participants

mentioned that they are not satisfied with only knowing the basic category name of visual attributes, or just vague comments, *e.g.*, "*Nice shape. This color looks great* ". They feel interested in the specific detailed descriptions and impressions on appearance. P5 mentioned his story that once he wanted to buy a guitar and was curious about what exactly was the color attribute 'gold'. "*Is it a shining one or a matte one? I want to know more details.* " For shape and size, participants always want to know how to compare the product with daily objects, *e.g.*, "*Does it fit on a tabletop?* ", "*Is it like the size of a banana?* ", which can help them better understand shape and size in the real world. Participants are also curious about if there are logos on the product.

To provide informative answers on basic attributes mentioned above, the key is to extract the descriptive and comparative details on appearance from the large amount of reviews. Particularly, clothes can have more nuanced design details to describe, such as sleeves, waist, neck, edition of clothing, graphic designs or patterns. These questions, *e.g.*, "*Do the buttons go all the way down? What kind of neck design does the shirt have?* ", are very category-specific and often require fashion knowledge to provide aesthetic descriptions, as P14 mentioned "*even my friends don't describe well* ", hence descriptions in common customer reviews hardly meet their expectations. In this paper, we don't focus on these subdivided fashion knowledge which requires large amount of annotated data (*e.g.* DeepFashion [48]), but focus on the common attributes shared by a wider range of products and possibly covered by customer reviews. The four attributes (color, shape, size, logo) are common across many products and can be concisely described in the alt-text compared to other aesthetic attributes (e.g., pattern, style).

2.5.2.2 Visual questions on high-level concepts

There are also high-level questions not directly related to basic visual attributes but can be inferred from the appearances in the image with commonsense, including usage, style, quality, texture and specific accessibility requirements, *e.g.*, , "*Does it look like it has good*

quality? How to open this bottle? Does the product have physical buttons? " Such properties are similar to the "signifier" concept proposed by Don Norman [3], which act as the indicator that can be interpreted meaningfully in the context of the social and physical world. These questions, usually framed as Visual Commonsense Reasoning, has been a challenging research topic in computer vision [49]. Yet currently it is not well-explored to address the need of inferring concepts and functions from product images for online shopping scenarios, as they require a large amount of annotated data from external knowledge sources beyond the shopping websites. However, the reviews commenting on the specific aspects (*e.g.* quality, physical buttons) can still indirectly provide insights on customers' opinions. Thus our expediency is treating these questions same as the non-visual questions to provide the relevant reviews.

2.5.3 Rules-Based Solution for Retrieving Informative Review Snippets

In this work, we mainly focused on addressing BLV users' questions on the common basic visual attributes, including color, logo, shape and size, which can be supported by augmenting keywords-based review searching with rule-based filtering (described below) to extract more informative snippets.

Compared to data-driven models, rule-based solutions can work well without the need of manually labelled data. Further, rules are explainable and can be validated by our targeted users. Rules are also modular—existing rules can be edited or removed and new rules can be added without affecting the others. Note that currently our rules are not designed to answer questions related to high-level concepts, which we consider as future work.

[3]https://sites.google.com/site/thedesignofeverydaythings/home/signifiers

2.5.3.1 Rules for Filtering Out Less Informative Reviews

We first established several simple rules to *filter out* review sentences that cannot be used to further the understanding of visual attributes with specific rules:

(i) Short: sentence of 3 words or fewer; *e.g.*, "*satisfied*", "*like the shape*", "*poor logo*";

(ii) Reference to the image: reviews that comment whether the actual product is consistent with the photo provided by the seller, *e.g.*, "*looks exactly as pictured*" "*not shown as picture*". These sentences mentioned by our participants as "*useless*" and "*annoying*" since they often occur in the reviews but contain little useful information for BLV customers.

2.5.3.2 Rules for Basic Visual Attributes

All reviews containing the query keywords are included as candidates and annotated using basic Part-Of-Speech tags [4], which are also known as word classes or lexical categories, *e.g.*, noun, adjective. Based on the aforementioned analysis, a group of three researchers iteratively established several rules for extracting informative reviews as follows.

Rule 1: Adjective + Keyword or Keyword + Verb + Adjective. The descriptive adjectives usually provide supplementary visual information. Examples include "*a shimmery purple*", "*crescent shape*", "*a very nice-looking etched logo*", "*squarish shape*", "*The bubblegum color is glossy and fun.*", etc. It is hard to enumerate all the possible descriptive words, instead we decided to obtain qualified snippets by differentiating those with evaluative adjectives.

The evaluative adjectives expressing subjective emotions can be vague hence not helpful to further understand the visual attributes. *e.g.*, "*nice shape* ", "*great color* ". We collected a list of such vague adjectives, including "*great* ", "*nice* ", *good*, "*bad* ", "*horrible* ", "*disappointed* " etc. These snippets are less informative but still provide subjective opinions, so we

[4]http://www.nltk.org/book/ch05.html

still keep them as candidates, but as the lowest priority among reviews following our rules.

Rule 2: 1st pronoun + ... + Keyword + ... + that/which/but/because. Rather than the simple sentences only containing expressions on attitude *e.g.*, "*I feel disappointed at the color.*", the sentences with clauses usually provide more detailed and useful information. Coordinating conjunction, *e.g.*, with "*but*" emphasizes the two statements contrasting or contradicting each other. Subordinating conjunctions, *e.g.*, with "*because*" often contain detailed reasons for customers' attitude towards the visual attribute. Examples include "*I love the color (Bubblegum), which I bought because it was the lowest cost for a color that would be difficult to misplace or forget while traveling*".

Rule 3: Comparative Expressions. Other than the common rules above, there exist some specific rules that work for particular visual attributes. (1) Keyword (shape) + "like/liked", *e.g.*, "*shaped like a Cola Bottle*". Comparing the shape of a product with a familiar daily object can be helpful for better understanding the shape; (2) Keyword (size) + "fit/for/of", *e.g.*, "*size fits in all cup holders*", shows reviews containing details on how the product fits in the settings are informative; (3) "than/more of" + Keyword (color), *e.g.*, "*it is a terra cotta than mocha*". Some products look different with the picture provided by the sellers. Sighted customers complaining about this kind of difference can also be informative for better understanding the actual appearance.

2.5.4 Generating Answers and Image Descriptions

Finally, we use the review snippets retrieved by the rules above to answer a BLV user's specific visual question or to compose a brief appearance description. As an answer to a BLV user's question, we provide original reviews retrieved by our rules divided into two lists—positive and negative—based on a sentiment analysis. We also provide a summary with the numbers of positive and negative reviews, and the top-3 informative review snippets from reviews across the two lists. To generate both the lists and the top-3 summary, we need to rank the review snippets. In particular, review snippets selected by Rule 1 with descriptive adjs,

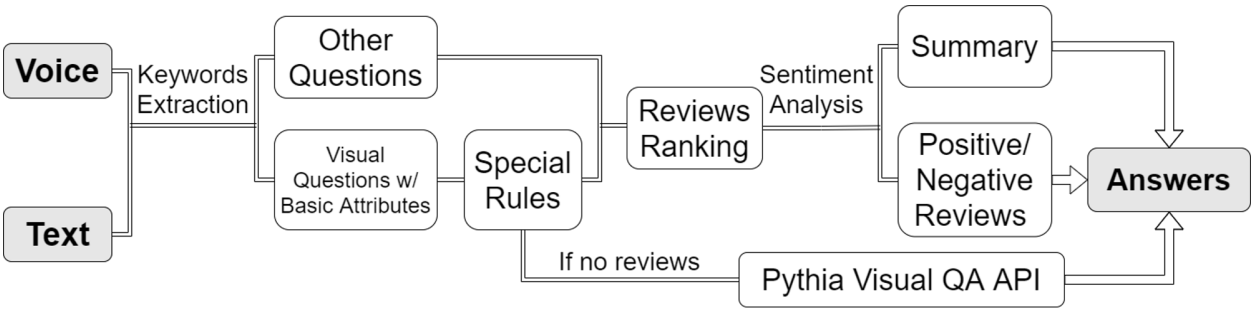


Figure 2.2: System overview of how Revamp responds to users' query. It first extract the keywords and process the query by its kind. After filtering out irrelevant reviews and reranking, Revamp uses the shortest candidate among the top-3 reviews to generate the visual description of a product and relevant answers to the query.

2, and 3 have higher priority than those by Rule 1 withe evaluative adjs; further, reviews of the same priority are ranked by their helpfulness (votes by other customers). The next level lower are reviews that contain the query keywords but do not meet any rules, within which they are ranked by relevance based on the concept of graph centrality following prior work [50]. For the brief appearance description, we select the shortest sentence in the top-3 candidates corresponding to each visual attribute (where each visual attribute is used as a keyword) since it is advised not to use long text for image alt-text. If users are interested in learning more details, they can move forward to ask specific questions on visual attributes.

2.6 Implementation of Revamp—An Information Retrieval System for Online Shopping

We present Revamp, an interactive information retrieval system to improve the shopping experience for BLV users. We implemented Revamp as a browser extension that works on Amazon. Revamp allows BLV users to interact with a simplified product page, access image description composed of relevant reviews, ask questions and receive a summary with original

reviews as response, as shown in Figure 2.2. On the front-end, Revamp uses Chrome API to automatically simplify the page and injects our search component to the current page, allowing user to send query and get a response from our back-end. On the back-end, Revamp extracts the keywords in user's query and matches the keywords with review snippets data. By ranking the filtered review snippets and applying sentimental classification, the back-end returns a review summary and a positive and a negative lists.

Data source. Our data source consists of two main parts: *(i)* Basic attributes (title, color, price) from information provided by sellers; *(ii)* Reviews (containing title, content, rating, helpfulness, date, and author); and *(iii)* Customer Q&A; We scraped data from Amazon product pages using the Python library BeautifulSoup combined with Selenium and stored it in `.csv` format. If a product browsed by a user already exists in our database, we will use it directly; otherwise, we will run the scraper to retrieve the data and save it in the database. Our Python Flask based back-end API server will use these data to generate responses to our front-end's requests. Before our user study and evaluation, we pre-scraped the product data to avoid the potential latency of retrieving it in real-time.

Web page simplification We provide an an alternate view of the original Amazon pages to improve the screen reader browsing experience. Our extension firstly rearranges the elements on the Amazon product page by manipulating the DOM tree We removed the irreverent information such as advertisements and promotions, and keep only the product details, reviews and images from the original page. Revamp would generate brief descriptions as the alt-text of the images using the aforementioned rules. Then we use Chrome APIs to inject our Revamp module into the product page. We strictly followed WAI-ARIA[5] standards to make sure all components have proper attributes and keyboard interactions.

Review Snippets Extraction. Revamp responds to both keyword queries such as

[5]WAI-ARIA stands for Web Accessibility Initiative – Accessible Rich Internet Applications. It is a specification published by the World Wide Web Consortium (W3C) that specifies how to increase the accessibility of web pages.

"*color* " as well as natural language questions such as "*Does the product have physical buttons?* ". For keyword queries on four main visual attributes, including color, logo, shape, size, we extracted the reviews with the pre-defined extended keywords based on our formulated rules. Specifically, for color, the extended keywords included 'color', basic color name such as 'blue' and the special name provided by sellers such as 'surf the web'. For other natural language questions, we used the Rapid Automatic Keyword Extraction (RAKE) library [6] for keyword extraction, Synset from WordNet [7] to get the groupings of synonymous keywords that express the same concept, then extracted the reviews with the certain and synonymous keywords.

Answer Generation. There are three types of information generated. *(i)* A supplementary textual description of product images; *(ii)* Two reviews lists: positive and negative; and *(iii)* A summary of reviews based on the two reviews lists. After we retrieved the ranked reviews, we divided them into positive and negative categories based on aspect-based sentimental classification by MonkeyLearn API. Finally, we extract top three reviews from each list to generate our summary using a pre-defined template. In case there are too few reviews to populate the answers, we leverage the existing computer vision technique to provide basic answers, *e.g.*, using Pythia [25] to answer the color, whether there is a logo on the image, and what is the shape. Users can either use text or voice input, when Revamp finishes generating answers, it will notify the user with a beep. Rather than automatically reading out the answers, we support users to use their own screen readers to scan the answers, which provides them with more freedom of control.

[6] https://github.com/csurfer/rake-nltk

[7] https://www.nltk.org/howto/wordnet.html

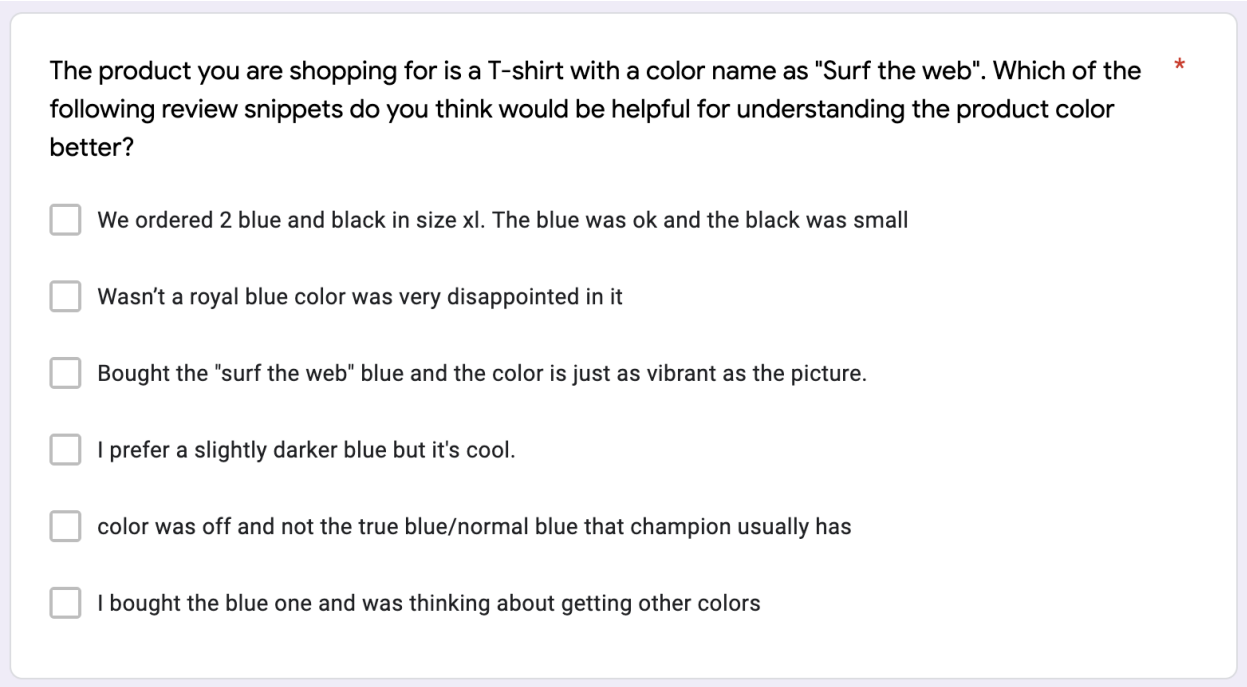


Figure 2.3: An example questionnaire for BLV users to vote for which one of the review snippets (retrieved by our rules or by default ranking of search result) best answers visual attributes questions.

2.7 Evaluation

We separated evaluation into two parts: rule evaluation and system evaluation. Rule evaluation is to evaluate the performance of rules-based solution for retrieving informative review snippets. System evaluation is to further evaluate how Revamp is integrated into BLV users' online shopping process. Modifying page structure and reviews-based interactions are non-separable, as they are coordinated to support a comprehensive information seeking flow. Hence we chose to evaluate Revamp as an integrated system.

2.7.1 Rule Evaluation

We first evaluated how our rules cover the informative review snippets for understanding appearance and how our solution can be generalized to different products.

2.7.1.1 Coverage of Informative Reviews for Understanding Appearances

We randomly selected eight products covering three aforementioned main categories from the Amazon best-seller list and extracted review snippets covering four main visual attributes (color, logo, shape and size) following the above rules. We then distributed voting questionnaires to eight BLV participants (P13 - P20). Given each product, a participants was asked to vote for the snippet that could help them best understand the visual attributes of that product: half of the snippets were generated by our proposed rules and the other half from the top-ranked answers in the existing Amazon 'Have a Question' section; all snippets were presented in a random order. Figure 2.3 shows an example question and review snippets. The snippets selected by our rules gained 85% of the 144 votes. Notice that we are not filtering out the remaining 15%. Review snippets which were not selected by our rules may also contain informative details, but seldom directly for describing the visual attributes, *e.g.*, a review "*We ordered 2 blue and black in size xl. The blue was ok and the black was small* " talking about color actually gave more details about the size differences between different colors. Rather than filtering out these reviews, we still show them but with lower ranking, while reviews that meet our rules and provide descriptive details such as "*color was off and not the true blue/normal blue that champion usually has* " will be assigned a higher priority.

2.7.1.2 Generalization to Various Products

To further validate the generalization of these rules in a more real-world setting, we downloaded web data of 45 selected products from Amazon's up-to-date top-seller ranking list (top 15 in each of the aforementioned three categories with similar products removed to maintain

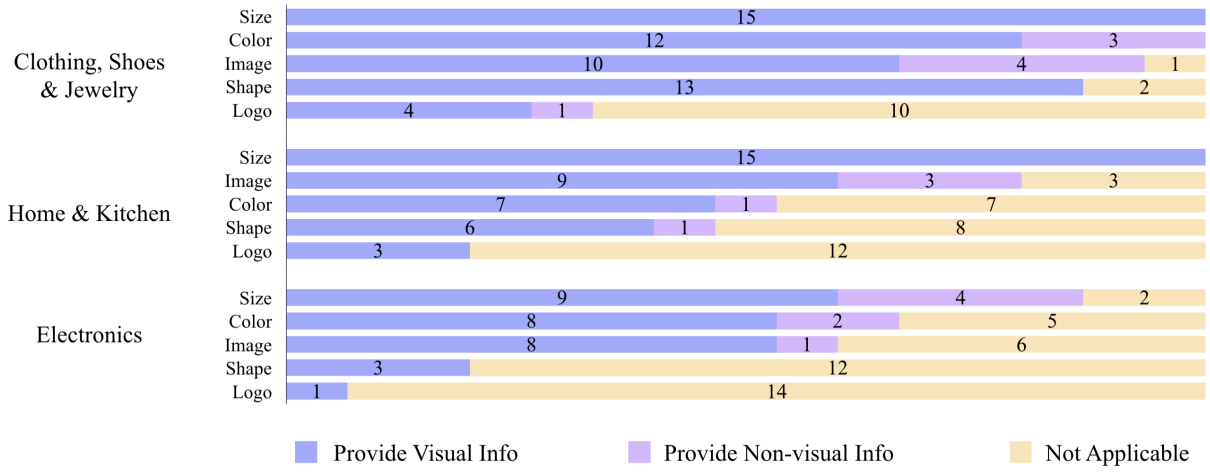


Figure 2.4: Performance of image description and visual questions on four basic attributes including shape, logo, size, color. Two sighted research team members rated the quality by three levels on 15 Amazon top-seller products in each of the three main categories. Results showed that our rules can directly provide visual info for 54 questions or image descriptions in Clothing, Shoes & Jewelry; 40 in Home & Kitchen; 29 in Electronics. We also included "Not Applicable" because of no relevant reviews or too few informative reviews.

variety). Two sighted research team members then browsed the product page and rated the quality of the generated alt-text and answers that describe four visual (color, logo, shape, size) in three levels: *not applicable, providing related non-visual information, providing direct visual information.* Results showed that our rules can provide informative results in most cases as long as there exist enough reviews, especially for the products whose appearances matter more in shopping decision, as shown in Figure 2.4. Not applicable is due to no reviews mentioned the attribute if the specific attribute does not occur (there is no logo on the product), or has no details to provide (*e.g.*, no reviews of a white bed sheet mentioned color details), or there exist too few informative reviews. In this case, Revamp uses Pythia Visual Question Asking [25] to provide back-up answers.

2.7.2 System Evaluation

To further understand how the system is integrated into BLV users' online shopping process, we evaluated Revamp with BLV consumers. We selected six representative products from the aforementioned three main categories including Water Bottle, Men's Tee, Women's Dress, Bed Sheet, Bluetooth Speaker, and Chair Cushion as the testing products. Three representative products, retrieved reviews by rules and generated image description are shown in Figure 2.5.

2.7.2.1 Participants, Design, and Procedure

We conducted interviews with the same eight blind participants (P13 - P20). Each user study lasted approximately 40 mins using Zoom after we received the IRB approval. Participants used their own laptops and screen readers and shared their feedback with the experimenters at the same time. We first asked the users to browse the products as they used to on original Amazon web pages without Revamp. The main task was to obtain information on four visual attributes (color, logo, shape, size) also other details they were curious about. We then introduced how Revamp worked, including how it simplified the original Amazon web page and what kinds of questions Revamp could answer, and asked the participants to repeat the same task using Revamp. Participants then filled out a survey with Likert-scale statements as shown in Table 2.1. Participants then further explained their reasons of giving each specific ratings. Each study was audio recorded and transcribed and participants' qualitative response was summarized by affinity diagramming.

2.7.2.2 Results

We observed the common interaction pattern with Revamp proceeded as follows: The participants first navigated through four elements (product name, image with alt-text, info provided by sellers and QA) on Revamp. They then asked questions on visual attributes

Table 2.1: We collected participants' subjective ratings on Revamp. The scale was 1 to 7, in which 7 = I strongly agree with this statement, 4 = neutral, and 1 = I strongly disagree with this statement. The value represents the average of ratings (SD). Data was analyzed using Wilcoxon test and a statistical significant difference ($p < 0.05$) is marked with $*$—all ratings of Revamp significantly outperformed the current practice.

Statements	**Revamp**	**Current**
It is easy and efficient for me to locate product-related information I need *	5.50 (1.07)	4.75 (0.89)
My questions are answered with informative answers *	5.13 (0.64)	4.38 (1.41)
I feel confident in understanding the appearance of the product *	5.88 (1.25)	3.00 (0.76)
I believe I can fully utilize the information from reviews *	5.75 (1.83)	5.13 (1.46)
Revamp can be a helpful alternative for shopping online independently when no sighted helper is available	6.50 (0.53)	-
I will use Revamp regularly in my daily life	6.00 (0.76)	-

and other aspects they were curious about (no predefined questions were provided). Since personal interests differ, they asked 2 ˜5 questions for each product, including both visual and non-visual questions. Besides the review summary, they also browsed the positive and negative review lists when needed.

Overall participants considered Revamp to be a helpful alternative for shopping online independently when no sighted helper is available and will use Revamp regularly in daily life. The comparative ratings are shown in Figure 2.6. Notice that one limitation of study is the bias of comparing a new system to an old one, participants would know which system was designed by the authors and can be influenced by this. Unfortunately, given the popularity of Amazon, participants would have known which one was ours even if we had intentionally de-identified the two systems. We further discuss participants' subjective comments below.

Category	Rule	Retrieved Reviews	Image Description by Revamp	Recognition by SeeingAI
Clothing, Shoes & Jewelry 	Rule 1 Rule 2 Rule 1	I believe is called "army green" is actually **more of** a **gorgeous kelly green** that is perfect **for** Christmas, **Green** is **vibrant** and true to color It was a little loose **but** overall I'm glad I went a **size** up Cute and **flattering shape**, Perfect for **hour glass shaped** ladies	Green is vibrant and true to color, it was a little loose but overall I'm glad I went a size up, Cute and flattering shape, no logo (from Pythia API)	Probably a green dress
Home & Kitchen 	Rule 3 Rule 1 Rule 1, Rule 3	**Spice** is like a **burnt orange**, if you ask me it's more salmon **than spice** **Oval shaped**, cushion **shape** is not **round**, **curvy (Concave) shape** It's a **giant size** cushion, **fits** well **for** the **size** compared to the frame (44" frame) and it is very plush	Spice is like a burnt orange, Oval shaped, It's a giant size cushion, no logo (from Pythia API)	Probably a person sitting on a leather couch
Electronics 	Rule 1, Rule 2 Rule 1 Rule 3	The **color** is **beautiful**, a **light aquamarine that** is very **mild** and **unoffensive**. the **aquamarine color** is **playful** and **fun**. has a clean look to it with its **squarish shape** almost about the **size of** a brick, Is the perfect **size for** so many uses	the aquamarine color is playful and fun, almost about the size of a brick, has a clean look to it with its squarish shape, has logo (from Pythia API)	Probably a close up of electronics

Figure 2.5: Examples of representative products in three main categories, showing the retrieved reviews, generated image description and comparing with the recognition results by SeeingAI, a common tool for BLV users to figure out what is in an image.

Providing supplementary information for better understanding the product appearance. With Revamp, participants' ratings on their understanding of product appearance were improved comparing with their current practice, owing to the supplementary image description and responses to visual questions. Having a supplementary description of the product image is a new experience to them: "*You know the Amazon doesn't provide image descriptions for the products. I will definitely use this add-on in my life.* " (P13) When using Amazon without Revamp, participants often just quickly skipped the image web component since no informative information is provided; with Revamp, participants tended to use the shortcut command of screen readers to directly jump to image to first gain an overall understanding of the product's appearance.

Participants felt surprised when Revamp provides useful answers on visual attributes, as mentioned by P19: "*It is really cool. It did answer my questions about product appearance.* " and even helps them learn about unfamiliar visual attributes, as shared by P13: "*Although I haven't seen colors before, I have a lot of fun in reading these descriptions of colors. For example, I don't know what is the 'Spice' color. It told me a review mentioned 'like a burnt orange', which is much more understandable.* " Participants also mentioned that descriptive details of reviews from the first-hand buyers sometimes felt more trustworthy than opinions of a friend or family member, as P15 commented "*Who can perform better than the customers who have bought the product themselves on describing the product? I really like the idea of using the reviews.* " Most participants agreed that they can better utilize the information from reviews. Only one participant gave a low rating of 2 because he personally preferred not to use any filter on the reviews in case something important is missed.

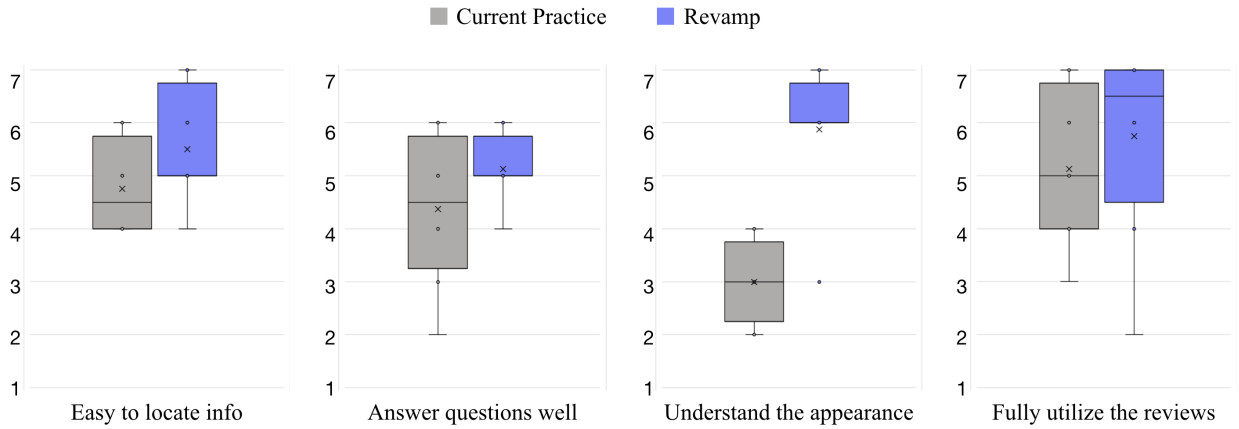


Figure 2.6: Subjective ratings on statements comparing current practice with experience of using Revamp. Revamp demonstrates a clear advantage in the experience of understanding the product appearance.

Enhancing interaction flow of information seeking. With Revamp, the information seeking experience of participants was improved comparing with their current practice,

owing to the reconstructed web page and better utilizing reviews. Revamp provided users with cleaner web page structure and is more user-friendly than the current Amazon page, as stated by P20: "*I don't need to worry of being stuck in useless information any more.* " Participants found the review summary and the review lists divided by positive and negative useful to access customers' opinions more efficiently and proactively. P14 mentioned that "*I like it that Revamp also keeps the original reviews accessible in the lists. After reading the summary, I can then make the decision to skip or look into the details of each relevant reviews in the list.*" Participants were more engaged in asking questions, as P19 commented "*Sometimes I could be inspired to ask more after I read the answers.* " Participants could choose to interact with the system using either voice commands or text input based on their preferences. Although participants in general liked the experience of screen readers augmented with voice input, they preferred to receive answers in text rather than speech: "*Using my screen reader to read out the answers is far more better than directly answering my questions by voice. I can adjust the speed and pause anytime.* " (P15)

2.8 discussion

Most web pages with vivid designs and a large amount of information are still not accessible enough to BLV users at present. We explored three aspects towards enhancing the information seeking experience on online shopping websites: *(i)* simplifying and reconstructing the web pages according to users' current task; *(ii)* providing coordinated experience of active query and passive reading to support flexible information seeking; *(iii)* leveraging relative text resources on the web page, such as reviews, to fill in the information gap. Besides, this work also inspired several exciting future directions as follows.

2.8.1 Working with Multiple Product Pages of Online Shopping

In this work, we focus on the task of understanding the product details, especially the appearance. Imagine another task: if the user has several product candidates in mind, we should also correspondingly meet a new information seeking need of comparing multiple products. To address this, future work can explore reconstructing the webpages to support efficient switching among products and answering comparative questions about multiple products. Currently, table is a common web page element for comparing different products and should be supported in future versions of Revamp. For example, the system should guide the user to navigate a table by informing them which products are being compared in terms what attributes and further be able to answer a user's question by looking up the table.

2.8.2 Working with A Broader Variety of Product Pages

Furthermore, future work can employ and test our methods on other product pages where there is often an overload of information inaccessible for BLV users to retrieve. For example, on Yelp, our methods can add description to popular dishes, whose images do not always convey all important information (*e.g.*, how large is the portion); on Trip advisor, images taken by fellow travellers can be further described using our methods by extracting relevant descriptions from others' comments (*e.g.*, whether a trail has shade); on Youtube, comments can be leveraged to generated video captions for eye-catching moments, which serve as a more vivid introduction before one decides whether to watch a given video.

2.8.3 Working with Supervised methods

Although the subjective comments were useful and convey vivid details, our participants pointed out that the generated appearance descriptions might be too subjective since it only contained one snippet in each visual aspect. However, if we follow prior work on filtering out

subjective comments [42], we will lose the vivid and descriptive details that are crucial to appearance understanding. As such, participants preferred the image description to remain as concise as possible and the description provided by Revamp would serve more as "*a first step or a clue.* " (P16) Also, it remains to be explored how we can better retrieve useful information from reviews on the high-level visual concepts inferred by image. In this work, we formulated hand-crafted rules to explore the possibility of leveraging reviews as an additional information source to fill the visual information gap. In the future, we can collect and label data to deploy supervised methods such as Knowledge Graph which could extract more information to better support queries not limited in the four visual aspects (color, logo, shape and size) and provide better summative appearance description.

2.8.4 Working with Better QA Experience

The question & answering experience of Revamp can be improved by providing prompting questions and involving more information sources. We observed that some participants had difficulty in formulating what questions to ask: P16 only asked about basic attributes, *e.g.*, color, quality. She explained that "*Sometimes I don't know how to ask a 'good' question. Maybe if I change my wording, I can get more answers.* " and suggested us to provide more pre-defined questions as prompts. Meanwhile, participants noticed that there still exists a gap between their expectations and the answer quality in other details beyond the four main visual attributes. "*Revamp can support descriptive answers for basic visual questions, but when I want to ask about some concrete (non-visual) details such as the dimensions of the bottle, it cannot answer me directly.*"(P19) In the future, we can involve more information source into Revamp, also leveraging Optical Character Recognition to extract the text information on the images provided by sellers.

2.8.5 Working in the Real-World Settings

Similar to other information retrieval system, the performance of Revamp highly depends on the quality of existing data. Currently, we leverage existing visual question answering API to provide a back-up answer for visual questions when there are too few reviews or no informative reviews. In the future, we can also involve human helpers in the system by sending the questions to crowd workers. Our rules for extracting informative reviews can also give insights on formulating guidelines of describing products for BLV users: *(i)* include descriptive details rather than vague opinions (Rule 1); *(ii)* if have to express personal attitudes, elaborate the reasons (Rule 2); *(iii)* compare with daily or common objects to help understand new concepts (Rule 3). Besides, participants also hoped to see Revamp's functions integrated into the mobile application with Alexa since they often try the mobile application of Amazon when faced with accessibility problems shopping on the web.

2.8.6 Limitations

There are several limitations of the system evaluation study. First, we did not counterbalance the conditions in the system evaluation. Using Revamp first would introduce a strong carry-over effect, as the information provided by Revamp is an augmentation of the baseline. As personal interests and product information differ, to compare intuitively how Revamp improves the experience, all participants first browsed the products as they used to on original Amazon web pages then on the modified web page of the same product by Revamp. Second, there exists the bias of comparing a new system to an old one when participants know which system was designed by the researchers [51, 52]. Third, only eight participants and a limited amount of products were involved in evaluation. This limitation can be addressed by a large-scale in-the-wild study, which we regard as future work beyond this initial study.

2.9 Conclusion

We present Revamp, a system that employs information retrieval techniques to meet the unique information seeking requirements of BLV consumers when shopping online. Our main contribution is a rule-based approach that leverages rich customer reviews to serve as image description and to answer BLV users' questions related to product appearances. Evaluations with eight BLV consumers showed that Revamp provides useful subjective information for understanding the product appearance and enhances the accessible information seeking experience of online shopping. Although Revamp could not provide answers for all of the products (*e.g.*, when there are too few reviews of a product), it can serve as an effective supplemental helper for BLV users to better access and understand a product before making a purchase decision.

CHAPTER 3

Nudging People to Use Visual Descriptive Language: An Application for Online Video Watching and Commenting

3.1 Introduction

The rapid growth in popularity of online videos has exacerbated the accessibility gap for blind and low vision (BLV) individuals who often lack access to the visual content in online videos [53]. To gain access to videos, BLV video consumers use audio descriptions (AD), or narration of the important visual details in a video that cannot be understood from audio alone[1]. While professionals craft audio descriptions for film and TV, such descriptions are expensive [54] and difficult to scale for online videos. Fully automated audio descriptions still do not meet content or accuracy expectations of BLV video consumers [55]. Other research attempts to bridge this gap by developing human-authoring tools to support novice describers such as volunteers and video content creators with writing high-quality descriptions efficiently [56, 4, 57, 58, 59]. Still, the sheer volume of online videos — with over 500 hours of new content uploaded every minute on YouTube alone[2] — makes it unlikely that the accessibility gap will be addressed through dedicated audio description efforts by a limited number of AD creators. In addition to assisting these AD creators who are self-motivated and

[1]https://www.access-board.gov/guidelines-and-standards/communications-and-it/about-the-ict-refresh/proposed-rule/

[2]https://www.globalmediainsight.com/blog/youtube-users-statistics/

aware of the task, an opportunity exists to expand the accessibility ecosystem by engaging sighted viewers in contributing to online video accessibility as they watch videos.

As videos become increasingly ubiquitous, so is video-based commenting. Human-authored reviews and comments can be used to describe visual media such as online product images and comics, but the effectiveness of such approaches requires the reviews and comments to specify visual details [10, 60]. Prior research revealed that YouTube comments referencing visuals can be difficult to interpret due to the lack of context [61]. For example, *"2:50 this kid is soooo funny xD"* only expresses an opinion about the visual entity *'kid'* without elaborating on visual details to clarify why the kid is funny. In our vision, sighted viewers can be nudged to contribute to video accessibility by anchoring and grounding their comments to specific visual contents, as a secondary task when they are watching the video. Such comments, although not equivalent to AD, can help BLV viewers understand inaccessible videos without AD.

The concept of nudging [9] has been widely applied by HCI researchers, to guide users towards desired behaviors in multiple contexts for social good, including health, sustainability and privacy. Fogg's Behavior Model [62] for persuasive design identifies three types of nudges for a target behavior to happen, including facilitators (ones that increase ability to pursue the behavior), sparks (ones that increase motivation for the pursue of the behavior), and signals (ones that merely indicate or remind of the behavior). We followed this behavior model to elicit interactive features that can be built into YouTube video web pages at watching and commenting time, including *a) visualizations of inaccessible video segments*, *b) hints on describing inaccessible segments*, *c) prompts of making comments accessible*, and *d) captions and comments as references*. As a proof-of-concept design probe, we developed CoSight, a Chrome extension to augment YouTube video pages with above features, as shown in Figure 4.1.

We studied CoSight's interactive features by inviting 48 sighted individuals who used YouTube on a regular basis to watch assigned videos covering four categories (popular sci-

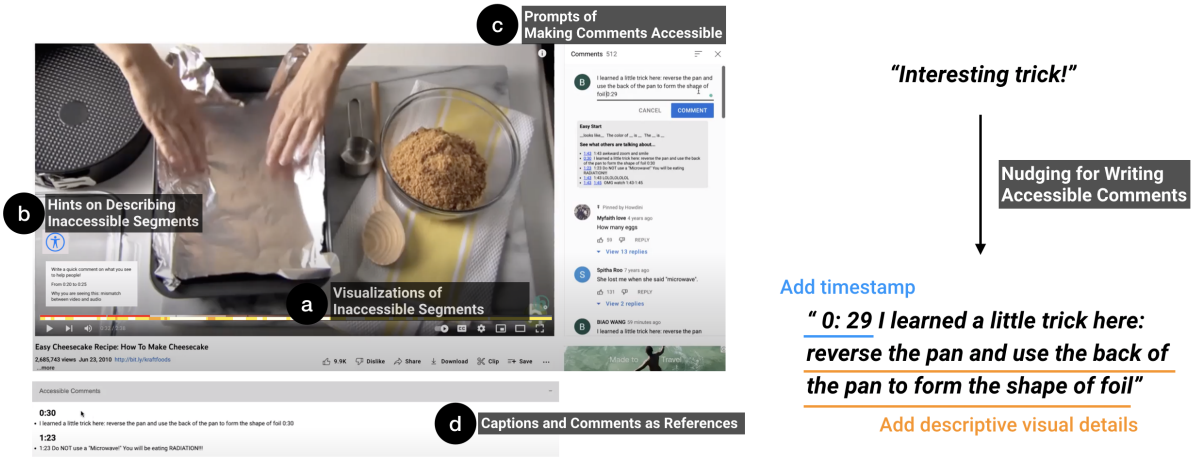


Figure 3.1: CoSight, a Chrome extension that augments YouTube video pages with interactive features to nudge the sighted viewers to contribute to video accessibility by anchoring and grounding their comments to specific visual contents of the videos. Blind and low vision users of CoSight can listen to these time-anchored descriptive comments of visuals as they play the videos.

ence, documentary, vlog, how-to) with CoSight. They decided if they would and how to write comments at their watching time, which resulted in 257 comments for 8 videos. Overall, analysis on the contents of comments showed that participants using CoSight included timestamps and descriptive visual details in their comments, with 94% of comments contained a timestamp, and 89% of the comments contained concrete visual details not covered by the speech of corresponding video segment. However, extending extra effort to add visual details while watching videos can still be mentally demanding for accessibility novices. To further understand if these comments can help BLV viewers understand visuals in videos, we invited four BLV individuals who use YouTube in a daily basis and have multi-year experience with AD, and four AD writers who completed professional training to provide their feedback on the comments content. We developed an accessibility mode of CoSight for BLV users to access the collect comments. BLV users can listen to time-anchored descriptive com-

ments by screen readers as they play the videos. Subjective feedback showed that accessible comments collected by CoSight contain visual details that were helpful for understanding the videos. Results showed that, while the quality of descriptions in such comments cannot be compared with professional AD authoring, these comments can provide BLV people with helpful visual information and also convey the vibes and emotions that are crucial to understand the videos. More importantly, this work highlights the potential of socio-technical systems like CoSight to not only enhance video accessibility for BLV individuals but also to foster a broader societal awareness and involvement in accessibility efforts. This approach has the potential to yield significant long-term social benefits by engaging sighted individuals in contributing to the accessibility ecosystem.

In summary, this paper contributes:

- CoSight as a design probe to explore accessibility nudges for online videos
- A user study with 48 sighted YouTube users writing video comments
- An evaluation with 4 blind people and 4 professional audio description writers

3.2 Related Work

Our work is informed by prior research on utilizing human-authored texts to enhance accessibility, authoring tools for writing descriptions, and nudging for social good.

3.2.1 Utilizing Human-Authored Texts to Enhance Accessibility

Crowdsourcing has been widely applied as an effective solution for tackling accessibility challenges in computing, particularly in providing descriptions to visual media such as videos [59], images [63, 7] and GIFs [64]. For example, YouDescribe [59] is a free crowdsourcing platform for YouTube that allows BLV people to submit their requests and supports sighted volunteers to script and record audio descriptions. Futhermore, organic crowdsourcing [65]

has also been applied in accessibility (*e.g.*, collecting sign language data [66]). This technique collects useful data from people while they take on tasks that they normally and intrinsically motivated to pursue, which can improved scalability compared with traditional crowdsourcing. Recent works have also suggested utilizing existing human-authored contents as a valuable source for information accessibility, such as online reviews [67, 68, 10] and comments [69, 60]. Revamp [10] extracts descriptive details from customer reviews of Amazon products to provide vivid visual descriptions covering attributes including color, size, shape and logo. Cocomix [60] extracts comments from viewers who are familiar with the story context to provide descriptions for relevant webtoon panels. These works underscore the advantage of utilizing human-authored contents to provide a vivid and contextual understanding, but also require such contents include descriptive visual details, which is uncommon.

While attempts have been made to automate the generation of audio descriptions for videos, they have revealed shortcomings such as context mismatch, lack of logistic, and missing detailed information [55, 70, 71, 72]. Previous research has explored utilizing time-anchored video comments for supporting online learning [73, 74, 75] and facilitating communication [76]. Additionally, studies have reported instances where BLV viewers sought to determine if a video is accessible through comments [11], and BLV vloggers utilized comments for interaction with their audience [77]. Yarmand *et al.* [61] found that visual entities are often referenced in YouTube comments. However, their analysis also revealed a deficiency in within-comment context, with many references only mentioning the entities but lacking concrete descriptive details, limiting their usefulness in helping BLV people to understand video contents. Therefore, this paper aims to explore the potential of nudging video audiences to contribute to accessibility by anchoring and grounding their comments to specific visual contents.

3.2.2 Authoring Tools for Writing Descriptions

Accessible videos for BLV people require a corresponding audio description, or a narration of the important visual details that cannot be understood from the main soundtrack alone [78, 79, 80]. Authoring audio description is a time-consuming process which requires knowledge of where, what and how to describe. Hiring professional audio describers to provide high-quality audio is costly and not readily-available [72], as estimated to cost $12 to $75 per video minute and the turnaround time from days to weeks [81]. While professional audio describers remain the gold standard, recent work suggested involving video creators, novices amateurs, and volunteers in the process to offer cost-effective audio description with assistive tools [82, 72, 83, 71, 84]. Kobayashi *et al.* [82] and Branje *et al.* [84] reported that it is feasible for amateur describers with little and no experience to generate effective audio descriptions supported by technology. Natalie *et al.* [83] developed a collaborative tool to support the sighted novice to author audio description with the commentary feedback from either a sighted or a blind reviewer. Technical solutions are developed to minimize the time and efforts of authoring descriptions by segmenting scenes [70], detecting visual content [70, 85, 86], identifying accessibility issues [53, 12], fitting human-authored descriptions into timeline [84, 4], and synthesizing text-to-speech for playback of the descriptions [87, 85, 88]. Prior work also pointed out various ways to improve description quality of people by making guidelines easier to understand with visualizations[89, 90, 91, 92], providing visual questions as prompts[93], showing high-quality examples [94, 95], and providing assessment [96, 58]. Different from assisting people who are either self-motivated or paid to write descriptions, this paper focuses on designing interface to nudge video audience who may be unaware of video inaccessibility problems to write comments with more descriptive visual details.

3.2.3 Nudging for Social Good

HCI Researchers have applied the concept of nudging [9] in multiple contexts including health, sustainability and privacy by designing systems to introduce subtle changes in the way information are presented with the goal of guiding users towards desired behaviors. Fogg's Behavior Model [62] suggests that for a target behavior to happen, a person must have sufficient motivation, sufficient ability, and an effective trigger. It identifies three types of triggers: *facilitators* (ones that increase ability to pursue the behavior), *sparks* (ones that increase motivation for the pursue of the behavior), and *signals* (ones that merely indicate or remind of the behavior). Caranan *et al.* [97] found 74 examples of nudging in HCI literature, identified 23 distinct mechanisms of nudging, and grouped them in 6 overall categories including facilitate, confront, deceive, social influence, fear, and reinforce. For example, confront nudges attempt to pause an unwanted action by instilling doubt. These works offer valuable insights for designing nudges to promote video accessibility. We are also inspired by how lightweight interventions can change people's behaviors on social platforms and doing good for others. Jahanbakhsh *et al.* [98] explored lightweight interventions such as checkboxes at posting time to reduce the sharing of misinformation on social media. Complex tasks can be break down into manageable, independently executable microtasks that can be distributed to the crowd and executed over time [99, 100]. Prior work has explored its applications in taxonomy-creation [101], writing [102, 103], conversation [104, 105] and scheduling meetings [106]. Especially for accessibility, Lasecki *et al.* [107] introduced a new approach in which groups of non-experts collectively caption speech in real-time to help deaf people. This paper seeks to extend the application into the unexplored scenario of enhancing video accessibility, by nudging the audience to contribute with lightweight and manageable efforts at watching and commenting time.

3.3 The Design of Nudges at Video Watching and Commenting Time

We define two levels of contribution that a sighted viewer may be able to make: *(i)* write a descriptive comment for an inaccessible segment; or *(ii)* make a comment more accessible by anchoring and grounding it to specific visual contents. We applied the Fogg's Behavior Model [62] to elicit the design of nudges in our application scenario: *sparks* to increase motivation of contributing to accessibility when watching videos; *facilitators* to increase ability of writing accessible comments with timestamps and descriptive visual details; *signals* to remind people to consider accessibility when writing comments. We presented the design of nudges at watching and commenting time, as shown in Figure 3.2.

As a person starts watching a video, *spark* is applied to increase their motivation of contributing to accessibility, by visualizing inaccessible video segments along the progress bar. If they are motivated to contribute, they can further trigger the hints on describing inaccessible segments, which increase their ability of contributing to accessibility as a *facilitator* by providing additional knowledge on what to describe. Then the person may choose to proceed to write a descriptive comment for the inaccessible video segment, or choose not to write, and continue watching.

As a person starts writing a comment, *signal* and *facilitator* are applied to remind them to consider accessibility and assist them with writing comments. If the person is motivated to write this comment by sharing their own opinions/feelings, showing prompts of writing an accessible comment is the last chance to remind them to consider accessibility as a *signal*. Then the person may choose to cover important visuals and clarify the referent to make the comment more accessible, or choose not to modify their original writing. If the person is motivated to write a descriptive comment by nudges at watching time, related captions and comments can help them frame their descriptive comment as a *facilitator*.

We further describe how the nudges work on a YouTube video page as follows.

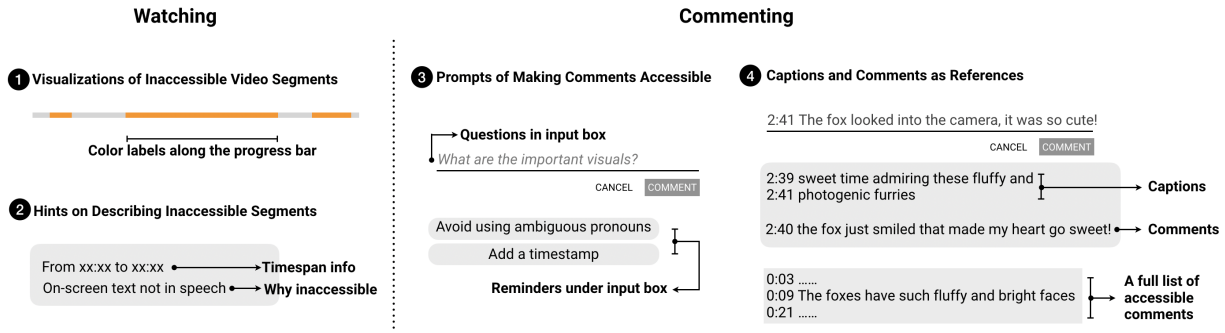


Figure 3.2: Features applying on video page, at watching time including 1) Visualizations of Inaccessible Video Segments; and 2) Hints on Describing Inaccessible Segments; at commenting time including 3) Prompts of Making Comments Accessible; and 4) Captions and Comments as References.

3.3.1 Visualizations of Inaccessible Video Segments

To motivate contribution, the video progress bar is enhanced with orange and yellow color labels under it, providing sighted viewers with a visual indication of the video's accessibility gaps. This overview not only highlights the areas where their input is most needed but also demonstrates the significance of their participation in improving accessibility. Since writing a comment for an inaccessible segment is a secondary task when watching videos, highlighting the top inaccessible segments as orange while labeling the second-tier inaccessible segments as yellow can help people prioritize their time and efforts on contributing to accessibility. As the video plays, these labels fade away when the mouse remains inactive, minimizing distractions during the watching experience. Instead, an accessibility icon appears briefly on the side of the screen when entering an inaccessible segment. Clicking this icon automatically positions the cursor in the comment input box and inserts the current video timestamp, streamlining the process of writing a descriptive comment, as shown in Figure 3.3.

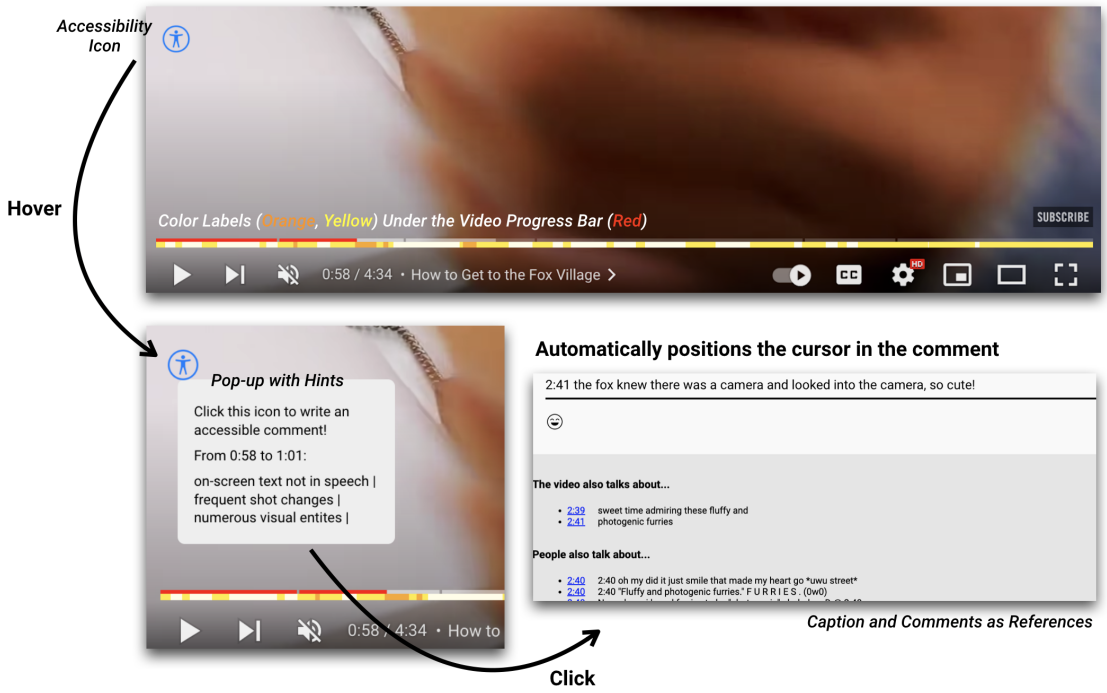


Figure 3.3: The interaction flow of writing a descriptive comment for an inaccessible segment: hovering on accessibility icon to trigger the pop-up card with hints on describing inaccessible segments, clicking on the icon to jump to input box with timestamp automatically added, then write the descriptions with related caption and comments as references.

3.3.2 Hints on Describing Inaccessible Segments

Pop-up cards can be triggered when hovering over the accessibility icon, containing the time range of the current inaccessible segment and brief words to explain why this segment may be inaccessible to provide knowledge on how to contribute. For example, *"On-screen texts not in speech"* indicates that clarifying the important on-screen texts that are not mentioned in the audio track may be helpful for BLV people to better understand the video content. The pop-ups will disappear automatically when not hovering over the accessibility icon to avoid blocking the video window at watch time. In addition, detailed guidance on what and how to describe is always be available to users to read at anytime, as shown in Appendix.

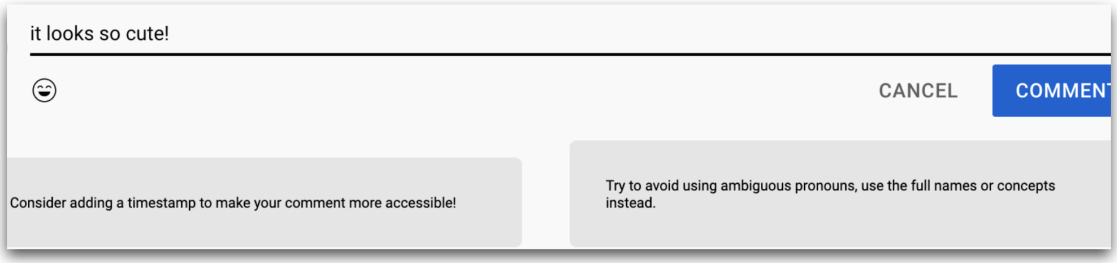


Figure 3.4: When the user write a comment to share their opinions, reminders on adding timestamps and clarifying ambiguous pronouns are shown based on the input content to nudge the user to anchor and ground their comment with specific visual detail.

3.3.3 Prompts of Making Comments Accessible

To facilitate the writing of comments, two prompt formats are implemented: 1) Light-grey question within the input box, *"What visuals not mentioned in the speech may be crucial to fully understand the video content?"*; 2) Banner reminders positioned below the input box, which are triggered if a sighted viewer doesn't include a timestamp with their comments or uses vague pronouns like "it," "this," and "that." These reminders prompt the user to specify the exact time and visual elements they are commenting, as illustrated in Figure 3.4. This feature accommodates sighted viewers with limited knowledge or motivation to write detailed descriptions, allowing them to contribute effectively with minimal effort to make their comments more accessible, even if they opt not to provide an extensive description for a particular inaccessible video segment.

3.3.4 Captions and Comments as References

Related captions and comments from other users will be displayed when users write their own comments to provide a sense of what information has already been contributed so they can add visual information not covered yet. Meanwhile, a full list of existing accessible comments will be displayed, and the comments people write with the tool will be shown in

this list as well. In addition, a priority list of existing accessible comments is also provided as reference. Such features may also motivates the user by creating the feeling of connectedness with the community by asynchronous cooperation and making people's contribution more visible.

3.4 Implementation of CoSight: A Chrome Extension On YouTube

We developed CoSight, a Chrome extension based on React, TypeScript, MongoDB, and craco npm package. CoSight augmented the Youtube video web pages by injecting the feature modules with Chrome APIs. Users can register a user account with email. After logging in, the beginner page of extension includes three tab pages: Intro, Manual, and Profile, which can be accessed by the Chrome menu: Extension - Options. The Intro page contains text guidance on why comments may help blind people and how to make comment helpful. The Manual page contains guidance on features provided by CoSight. The Profile page shows the historic video comments the users wrote with CoSight.

3.4.1 Identifying and Explaining Inaccessibility

To identify inaccessible video segments and provide explanations of why a video segment is not accessible, CoSight combined two computational pipelines from prior work on identifying video accessibility issues [11, 56].

CrossA11y [56] segments the auditory and visual track of the video and identifies asymmetries between auditory and visual tracks using cross-modal grounding analysis. As a black box method based on the MIL-NCE model [108] and the state-of-the-art MultiModal Versatile (MMV) networks [109], it can not provide concrete reasons on what is inaccessible, but it can rank inaccessibility of video segments by measuring the asymmetries as visual accessibility scores (0.0 - 1.0). For example, if a video segment contains many visual objects but not mentioned corresponding in audio tracks, it will have a low visual accessibility

score. The other work [11] defines video accessibility metrics to explain why one video segment is considered as accessible or inaccessible. For example, if a segment has frequent shot changes, it may be inaccessible and needs more description about the changing scenes. For people without accessibility knowledge, providing concrete explanations on inaccessible and suggestions on what to describe can help them authoring comments.

CoSight first utilized the CrossA11y pipeline to separate the video into segments and assign each segment an accessibility score, and select the top 5 inaccessible segments by scores, which are labelled as orange color along the progress bar. CoSight then followed the accessibility metrics to locate second-tier inaccessible segments which are labelled as yellow color along the progress bar, and provide hints for describing the inaccessible segments. A full list of explanations provided to users are shown in Appendix.

One limitation of the current implementation is the time cost, which requires around 2 mins processing for one 5-min video. The computational pipelines can be replaced by future available advanced techniques to support real-time processing of any YouTube video, while we focus on designing the front-end nudges in this paper.

3.4.2 Analyzing and Storing Comments

CoSight used NPM package *keyword-extractor*[3] to extract keywords of each caption sentence, same for each comment. CoSight leveraged these keywords and timestamp info to provide related caption and comments as users input. To select comments included in the accessible comment list, we filtered out comments without timestamps and have low semantic textual similarity with corresponding captions leveraging the SentenceTransformers model from Hugging Face API [4]. All the data objects are stored in the Mongo database.

[3]https://www.npmjs.com/package/keyword-extractor

[4]https://huggingface.co/tasks/sentence-similarity

3.5 Study with Sighted Viewers: Does CoSight Nudge People to Write Accessible Comments?

We conducted an evaluation study with 48 sighted YouTube users writing comments with CoSight to understand the usability of CoSight interface for sighted viewer, and the quality of Comments written with CoSight.

3.5.1 Methods

We recruited 48 participants (Female: 23, Male: 25, average age: 37) from Prolific who used YouTube at least once a month. We compensated each participant $12 for the study. Participants were informed that their compensation did not depend on the quantity or quality of their comments, and that commenting was entirely voluntary. Participants downloaded the zip file of our extension demo and installed to Chrome. They first read instructions on how to write comments for BLV video viewers using the extension. Then, they watched and optionally described two randomly selected videos. Finally, users filled in an online questionnaire to rate the usability of CoSight and answer open-end questions on reflecting helpful features and giving feedback on future improvements. Participants were divided into two groups randomly to use CoSight in two days separately, in this way the first group of participants only had access to existing comments, but the second group of participants also had access to comments written by the previous group.

3.5.1.1 Videos and Data Analysis

Eight videos, covering four categories including *popular science*, *documentary*, *vlog*, and *how-to*, were randomly selected from YouDescribe, a platform where BLV people can request audio descriptions for YouTube videos. Time length of videos range from 2 - 10 mins. Two researchers followed the thematic analysis methods [110] to open-code the participant-written

video comments and the subjective feedback independently. Then, the researchers reviewed each other's codes and discussed the codes to reach a consensus.

3.5.2 The Usability of CoSight Interface for Sighted Viewers

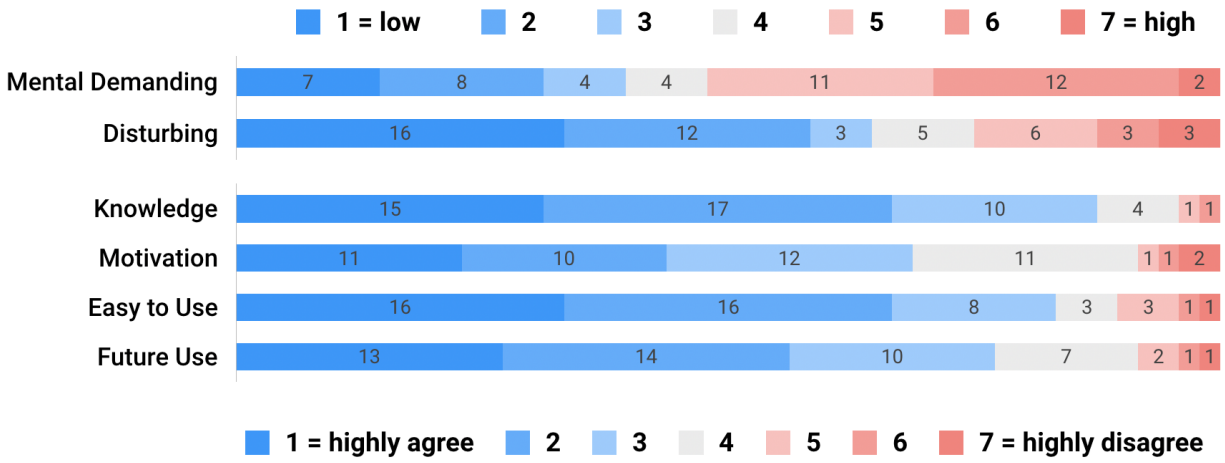


Figure 3.5: Participants reported ratings from 1 to 7on the user experience of CoSight, including the following statements: 1) How mentally demanding was writing accessible comments? (1 = low, 7 = high) 2) How distracting of writing accessible comments to the video watching experience? (1 = low, 7 = high) 3) After using this tool, I understand the concept of video accessibility. (1 = highly agree, 7 = high disagree) 4) This tool motivates me to contribute to the accessibility of online videos. 5) This tool makes it easier for me when writing accessible comments. 6) If these features are built in Youtube, I will use this tool to make my comments more accessible when watching videos in the future.

Overall, participants stated they would like to use CoSight in the future (37/48). CoSight empowered users with the knowledge to understand the concept of accessibility (42/48) and motivated them to contribute (33/48) by promoting awareness of accessibility, as *"it can help BLV people in ways that people don't normally think about"* and *"it creates a positive environment that promotes collaboration among the audience"*.

3.5.2.1 Writing Accessible Comments Intuitively Fit into the Interaction Flow

Most participants (40/48) agreed that CoSight makes contributing to accessibility easier for them, because the interaction flow was *"quite intuitive and simple"*, and the nudges *"fit smoothly with the YouTube video page."* Specifically, 7 participants mentioned the just-in-time pop-ups were helpful in reminding them of visuals they can describe when writing comments; 5 participants mentioned the shortcut of clicking on accessibility icon reducing their workload by automatically jumping to comment function and inputting the timestamp. Showing related accessible comments written by others is mentioned as helpful by 11 out of 24 participants who did the second-round study, as it saves their efforts for valuable contributions by *"preventing writing redundant comments"*. In contrast, first-round participants didn't find this feature useful since they didn't have access to accessible comments written by other participants with CoSight, but only the regular comments already existed. 4 participants reported feeling distracted from watching videos especially when they had to manually rewind to make sure they wrote an informative comment for a video segment containing various visual details. To optimize the flow, they suggested to support adding a *'bookmark'* or *'shortcut'* for one segment to revisit and rewind conveniently when they hoped to finish watching the video first.

3.5.2.2 Writing Accessible Comments as a Mentally Demanding Task

Still, about half of participants (25/48) regarded writing accessible comments as mentally demanding. Most of them (19/25) mentioned the mental efforts mainly came from the mismatch between the motivation of commenting and the task of describing. Based on participants' past experiences, comments have always been used to sharing opinions rather than describing visual details. As they shared, *"It's difficult because I don't feel that an natural comment is very descriptive."*, and *"I don't normally write comments with descriptions, so this was a bit mentally taxing."* 6 participants reported that applying text instructions on

writing accessible comments (as shown in Appendix) is mental demanding to them as first-time learners. This one-time learning step has long-term benefits, as one participant shared, *"it was somewhat intimidating when not familiar with the concept at the beginning but much better when I get used to it, and I will be able to contribute better comments for blind people."*

Compared with providing a description in comments, modifying a comment to make it more accessible is much easier to do, as participants shared: *"I like that there are just basic tweaks to ordinary comments to make them accessible. Like, if commenting on a particular frame, begin with a time stamp and avoid vague language or ambiguous pronouns."* and *"It is an easier starting point for someone without much knowledge on accessibility and this would make more sighted people want to contribute."*

3.5.3 The Content of Comments Written with CoSight

Our analysis demonstrated that participants using CoSight included timestamps and concrete visual details in their comments. In total, participants wrote 257 comments for 8 videos, with the length from 13 characters to 438 characters (average: 93 characters). 4 participants only wrote one comment for one out of two videos they watched. 19 participants wrote two comments, one comment for each video they watched, and 7 participants wrote more than five comments per video. The histogram of counts of comments written per participant is shown as figure 3.6. Video titles and accessible comment examples are provided in the Appendix.

3.5.3.1 Comments with a Timestamp.

As shown in Figure 3.7, 94% (241/257) of comments contained a timestamp, and 93% (223/241) of comments with a timestamp were aligned with segments labelled as inaccessible.

Among all participants, only 4 participants wrote comments without a timestamp. 2 of them reported that they intended to add timestamps but encountered a technical issue of

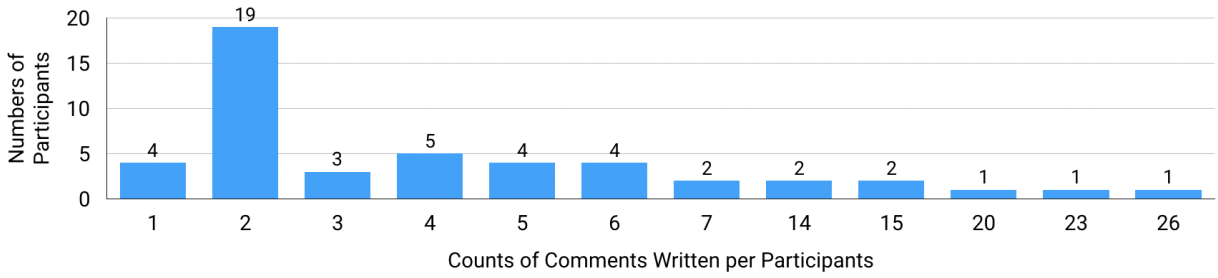


Figure 3.6: The histogram of counts of comments written per participant. Each participant watched two videos.

extension not responding accidentally. The remaining comments without a timestamp were general opinions on the whole video, such as "*I like the detail he is explaining in regards to the nervous system.* " 93% (223/241) of comments with a timestamp aligned with the inaccessible segments labelled by CoSight, while the inaccessible segments at the end of a video were likely to be neglected. For example, at the end of an educational video, there are many links and thumbnails of other related contents not mentioned in the narrations, and labelled as an inaccessible segment. But no participants choose to write a comment for this.

Participants still wrote comments for unlabeled segments when they found visual entities necessary to describe as the automated methods may not identify every inaccessible segment. For example, when the narrations introducing the fox village with the scenes walking around, a participant provided an additional visual detail even it was not labelled as inaccessible: "*01:25 a close up on a statue of a red fox* ". Some participants wrote comments when the segment was accessible, but they only wanted to share an opinion or feeling. For example, when watching the scene of a lot of cats on an island, a participant wrote "*0:19 I would love to visit the cat island!* ". The narrations already mentioned the cats, and the participant didn't provide any more visual details.

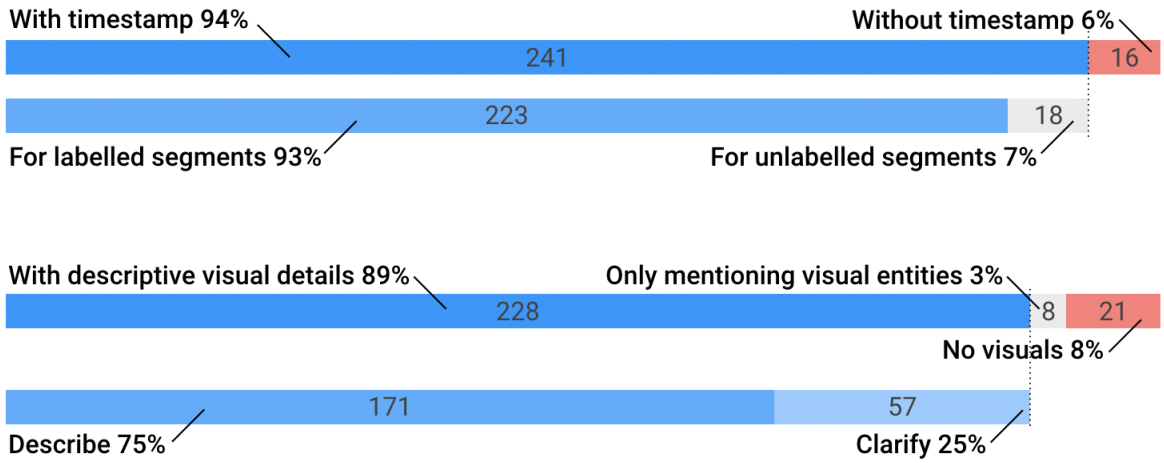


Figure 3.7: Analysis on the content of 257 comments showed that participants using CoSight included timestamps and descriptive visual details in their comments. Participants mainly wrote comments to describe labeled inaccessible segments, while they also wrote comments to share opinions with clarified visual referents.

3.5.3.2 Comments with descriptive visual details.

As shown in Figure 3.7, 89% (228/257) of the total comments contain concrete visual details not covered by the speech of corresponding video segment, 3% (8/257) of the total comments mentioned visual entities but didn't covered any concrete descriptive information, 8% (21/257) of the total comments purely expressed participants' opinions without mentioning visuals.

Among the 228 comments with concrete visual details, 75% (171/228) were written in the format of describing an inaccessible video segment, such as "*0:09 This shows a sunny, outdoor open area with trees, paths, and buildings in the background. There is a group of students standing talking to each other and we see two students walking together on the path.* " providing a detailed and objective description of a video scene without narrations. The remaining 25% (57/228) were in the format of sharing feelings but making an additional

effort to make their comments more accessible by clarifying the visuals, such as "*3:31 Oh my! There's one white fox in all of the reddish ones!* ", and "*2:07 The cat is so cute chewing on the brush oh my gosh* " providing more visual contexts when expressing opinions with adjectives and interjections. These comments provide descriptive details to scenes (78/228), concepts (52/228), actions (44/228), objects (30/228), animations (12/228), and texts (12/228) shown in the videos.

3% (8/257) of the total comments mentioned visual entities but didn't covered any concrete descriptive information, such as "*3:21 various images to accompany the discussion* ", and "*4:54 the cells look very cute* ". The former one indicated an attempt from the user trying to mention visuals but failing to provide informative details, and the latter one indicated a sharing of opinion from the user without making efforts in providing more descriptive details. There also existed exceptions: sometimes mentioning the visual object was enough since it is not necessary to describe how the object looks in detail if it was not directly related to the video theme, or the detail was already covered in the audio track and can be redundant if described again. 8% (21/257) of the total comments purely expressed participants' opinions without mentioning visuals, such as "*Seattle is a beautiful city* ".

3.6 Interviews with Experts: Do Comments Written with CoSight Really Work?

With this study, we hope to have a deeper understanding of: From the perspective of BLV viewers, do these comments written with CoSight help with understanding video contents? From the perspective of professional AD writers, can these comments work equivalently as AD? We conducted expert interviews with 4 BLV people and 4 AD writers to understand the effectiveness and limitations of this middle-ground solution between regular comment authoring and professional AD authoring.

3.6.1 Method

3.6.1.1 Demo for BLV Viewers

To demostrate for BLV people, we developed an accessibility mode of CoSight interface. As the BLV viewer plays a YouTube video, a beep notification sound will be delivered at the end of a video segment when there are available comments for this segment. The video will be paused automatically as default, while user can also change the setting as not pausing automatically. BLV user can then press Shift to navigate through and read out the available comments, and press Space to exit the comments reading and resume video playing, which is aligned with the original shortcuts of YouTube. BLV users are also supported to press Page Up or Page Down to provide feedback on if the comment contains useful visual details. We followed Web Accessibility Initiative - Accessible Rich Internet Applications (WAI-ARIA) standards to make sure web components of CoSight, such as the list of accessible comments, have proper attributes and keyboard interactions to be compatible with screen readers.

3.6.1.2 Interviews with 4 BLV people

Participants (U1-U4) were invited from an online discussion group of Audio Description and compensated $20 for the one-hour interview. Demographic information is shown in in Table 3.1. The participants walked through the interface of CoSight with the experimenter and played four videos, one in each category. Then the experimenter asked participants semi-structured interview questions in two themes: the quality of the comments, and their concerns and expectations on future improvements.

3.6.1.3 Study with 4 AD professional writers

Participants (W1-W4) were invited from a training program of professional AD writers and compensated $20 for the one-hour interview. AD writers walked through the interface of

Table 3.1: Demographics of blind participants.

ID	Age/Gender	Visually Impairment	Audio Description Experience
U1	37/M	Totally Blind	Use it in daily life
U2	28/F	Totally Blind	Use it in daily life
U3	34/F	Low Vision	Use it in daily life and also work on Professional training of AD writers
U4	41/M	Totally Blind	Use it in daily life

CoSight with the experimenter, tried out the features for sighted viewers when watching two videos of different categories, and gave feedback on the quality of representative collected comments.

3.6.1.4 Data Analysis

With user consent, the interviews conducted on Zoom were recorded and transcribed. One researcher followed the thematic analysis methods [110] to open-code the recorded data and the other researcher reviewed the codes. The researchers met to discuss the themes and reach consensus.

3.6.2 Results

3.6.2.1 Comments with Descriptive Details Help, But Are Not Equivalent to AD

When sharing their prior experience, all BLV users mentioned the comments are often very subjective and don't offer detailed information for understanding the visuals.

> *It's usually people disagreeing or agreeing something. Or it'll be something like 'Hey, did you see the thing at 2:55? That was really cool!' Well, that's not helpful*

without telling what exactly they're saying. (U1)

After watching videos with CoSight, BLV participants unanimously gave a 4 out of 5 on the agreement (5 = strongly agree) to the statement *"With the comments provided, I can better understand the visual content."* They agreed that most comments with timestamps and concrete visual details were helpful. For example, a comment *'1:29 Green algae swirls in the wake created by a tiger swimming across the screen.'* describing a dynamic scene without speech, was regarded as helpful by U3 since it provided visual information that she would not have had. Still, comments that purely expressing opinions without visual information were distracting to U1, like one comment of the same video, saying *'It is crazy how Tigers prefer to be in the water contrary to most cats.'*.

> *The majority of the comments were pretty helpful as far as being descriptive. But I don't want to make a blanket statement, say, 'definitely all helpful'. I would say no to opinionated comments.* (U1)

While BLV participants gave credit to comments providing visual details unknown to them, AD professional writers have more stringent standards on the quality of descriptions due to their training experiences. Overall, all AD writers agreed that most of collected comments cannot compare with the quality of professional AD authoring, as they mentioned two common problems of novice writers without training, either *"including not enough details"* or *"over-describing"*, also applied to these comments. For example, when writing for a cat bathing scene, a collected comments just briefly touch on "*the man is rubbing the cat now* " while W1 demonstrated that she would provide more descriptions on actions such as "*hands rotate around to massage and move up and down the cats body* ".

However, AD writers did not believe that it is necessary to compete the quality of comments with AD.

> *I always spend a lot of time revising my descriptions, given the nature of such*

comments are written quickly just at the time of watching, at least it works well on providing more visual details. (W1)

Although comments are not doing it equivalently as AD, it doesn't mean that it worth nothing. These comments can still promote video accessibility. (W2)

3.6.2.2 Comments with Subjective Opinions As Double-Edged Sword

BLV participants' attitudes toward the subjective comments differ. While U1 and U3 tend to not read comments with any subjective opinions when looking for descriptive information, U2 and U4 accepted reading descriptive comments with personal opinions involved, which can be fun and help them better understanding of video contents.

Comments definitely helped with the context. Obviously, they can be very subjective, but it helps give me an idea like about little nuances that I would otherwise miss. (U2)

Occasionally, comments with opinions can be useful when it pointed out some errors in the visual frame, such as "*1:10 Every part of the tongue detects sweetness. There's an inaccurate tongue map diagram.* " Meanwhile, participants understood it is within expectations that everyone writes comments in their own language style and not everyone can write a professional description.

You don't have to be a professional describer. It can be beneficial when humans helping out humans by telling me a few things that I'm missing. If it's not perfect, it is okay, we're still getting the info. (U4)

Interestingly, echoing with the mismatch between commenting and describing from sighted viewers, all BLV participants mentioned "*nomenclature* ", the wording used in the interface, can be important to nudge the sighted audience to share more objective descriptions than subjective feelings.

> *The term 'comment' may naturally let other people think 'Oh, they want my opinion of it' but the term description may let people feel 'you are asking me to relay information to you'.* (U1)

All AD writers mentioned that there exists a grey area for subjective and objective contents in comments, which can be hard for general audience to distinguish and decide what to include and what to remove. While personal opinions are definitely not accepted in professional AD, they feel a mix of opinions and descriptions is acceptable in comments, as long as BLV people are aware of that these comments are not written by professionals.

> *Different from audio descriptions, I think comments are just fun! Being totally objective will lose the fun. The vibes and emotions are also important information to convey.* (W2)

3.6.2.3 Social Benefits and Concerns

All four BLV participants gave a 5 out of 5 to the statement *"I will use this kind of technique or extension in the future."* and commented CoSight as 1a really dynamite service, 2a neat tool, and 3a good middle-ground solution of raising awareness. Different from YouDescribe, a separate crowdsourcing platform that volunteers add audio descriptions based on requests from visually impaired people, the spirit of CoSight is to take proactive actions as a sighted YouTube user without expertise in accessibility.

> *I think it could be very useful to to blind audience. This reminds me of YouDescribe. But this [CoSight] is baked into YouTube in a more primitive way.* (U1)

> *I would definitely enjoy this quite a lot because of the intimidating way it approached to improve awareness...the idea is like, we're not throwing out like 508 compliance, ADA standard, it's not like scary and a ton of extra work. Once people understand this, it's not super complicated to contribute.* (U3)

AD writers also see the value of CoSight not only in improving accessibility of specific videos for BLV people, but also promoting awareness of contributing to accessibility among sighted people, which has long-term social benefits.

> *It may not be most accurate or most beneficial in providing descriptions from the standard of AD, but it does raise consciousness and awareness to engage more people in the society in video accessibility.* (W4)

The main concerns mentioned by participants are around bias and abuse. 3We all have a description bias. If I say something is large, my large might be your small. It is necessary to educate people to try to keep their biases out when writing descriptions. While in the current development, CoSight presented human-authoring comments without reviewing, AD writers worry about deploying it on a large scale. 1Now we are testing it in a really controlled scenario. But when you put it in the wild, it's going to be unordered chaos. People may abuse the system and data.

3.7 Discussion

Moving beyond assisting people with specific expertise in accessibility, this work opened up new opportunities of engaging people without prior expertise to contribute to accessibility with manageable efforts integrated into their original interactions. As a design probe, this work points to a future where social and crowd intelligence can be central contributions to information accessibility, also elicits discussion on challenges and potentials of deploying such idea in the wild with collaboration of AI techniques.

3.7.1 Nudging People with Limited Knowledge with Scaffolding

Scaffolding is a frequently used strategy to facilitate learning by enabling learners to accomplish tasks that would be beyond their capacity if unassisted [111]. In our study with sighted

viewers, participants first spent time on reading through the instructions on writing accessible comments. In the wild, it may be difficult to require a random audience of one video to read instructions before start watching a video and writing a comment. Hence, breaking down the knowledge into scaffolds can help gradually involve people without awareness and motivation to contribute to accessibility, make them first care more about accessibility, then develop expertise step by step. For example, the pop-ups can be further leveraged as an educating tool, e.g., delivering good-quality descriptive comments as people watch videos, hence help them understand what comments are accessible in the practice.

3.7.2 Nudging People with Limited Motivation with Gamification

Gamification is defined in HCI as the use of game design elements in non-game contexts [112], which applies game elements such as challenges, levels, avatars, points, achievements, stories, and leaderboards into applications which their main purpose goes beyond pure entertainment. While CoSight serves as a useful system to nudge people without awareness and expertise to contribute, there are potentials to make contribution fun and motivate people to do it voluntarily with more engagement. For example, at the time one sighted viewer finishes writing an accessible comments, a celebration animation can be shown; sighted viewers can receive the upvotes and feedback from BLV viewers; and a leaderboard can be provided to reward people who contributes more to one inaccessible video.

3.7.3 Potential and Concerns of Involving AI in the Loop

Researchers have explored different approaches of generating descriptions of visual media, including fully human-authored (e.g., [5]), fully automated (e.g., [6]) and hybrid techniques of computer vision algorithms co-authoring with humans (e.g., [7]). Most efforts in mainstream solutions have concentrated on achieving acceptable results based on the visual media itself, but it is becoming evident that researchers need to question the one-size-fits-all so-

lution without considering different contexts [8]. Large language models can be applied to rephrase the collected comments into curated AD language style, or generate descriptions with multi-modal meta-data, including video itself, captions, human-authoring comments. Meanwhile, involving AI for summarizing the comments may lose the creativity and diversity of human-authored contents, also aggravate prejudice if a similar expression is used in multiple comments even it is harmful. Although there is potential to reduce the authoring efforts by supporting human to modify a generate result, researchers should also be cautious when involving AI in nudging for accessibility, as prior research on alt text [113] authoring has revealed that representing AI-generated info to assist content authoring may decrease the quality.

3.7.4 Values and Challenges of Deploying in the Wild

The idea of nudging for accessibility can be generalized to various online social platforms, such as nudging people to add more descriptive visual details in their reviews for online product, or in their comments for other visual media such as comics. It is worth mentioning that the values of such a system is two-fold, *(i)* on a practical level, it promotes the accessibility of online visual media, and *(ii)* in the long-term, it raises social awareness of accessibility and builds an accessibility ecosystem by engaging more people in the society to contribute. BLV people and sighted people can work together to co-create an accessible household [114] and workspace [115], also online community. Reflecting on the history of YouTube, it used to allow community captions, which allowed viewers to add subtitles to videos. This feature was discontinued on 2020, citing low usage and abuse, which also indicates that desired behaviours will need to be nudged in certain ways. While we may address the algorithmic challenges of processing data with bias by leveraging technical solutions, policymakers and platforms should make a continued effort on developing socio-technical solutions to promote social awareness of accessibility and avoid abuse [116, 117].

3.7.5 Limitation and Future Work

As it is still challenging to leverage state-of-the-art AI algorithms to process videos of varying categories and lengths in real-time, we conducted the study with sighted viewers on a selection of representative videos which were pre-processed for study users to access. Participants were informed that their compensation did not depend on the quantity or quality of their comments, and that commenting was entirely voluntary. Despite such efforts, this study could still not fully capture the nuances of behaviors in real-world setting. Future work could benefit from a long-term study that compare the behaviors with or without delivering nudges to participants when they watch videos typical of their usual viewing habits, covering a wider variety of video categories and lengths. Such a study would provide deeper insights but would require significant computational resources and joint efforts from the video platforms. As discussed in prior work [97, 118], nudges work best under uncertainty that is when individuals do not have strong preferences. The ineffectiveness of nudges when they does not work for particular people and in particular contexts, can be seen in a positive light since it may imply that the nudge preserves individuals' freedom of choice.

CHAPTER 4

Creating Symbiotic Information Seeking Experience: An Application for Online Research Paper Reading

4.1 Introduction

Figures in scientific papers serve as vital components to efficiently communicate crucial information, *e.g.*, teaser figure often provide a holistic view of research contributions, graphs and charts convey important results. However, a large majority of scientific paper PDFs lack usable alt text for figures [119], negatively impacting the reading experience of BLV people even if the rest of text in a paper were accessible. Recent work [120] has investigated the content of author-written alt texts and whether they convey adequate information about figures to BLV readers, showing that most alt text only contain basic information about what is the figure, far fewer contain descriptive information discussing what insights that figure try to convey. Attempts in automatically generating alt text on web and social media platforms do not apply as well to scientific figures, since most current models are trained on datasets that primarily composed of natural images. Hence, many established image understanding models cannot be directly translated to the scientific domain. Hybrid crowdsourcing solutions that combine human expertise with machine intelligence have emerged as promising alternatives [7, 121, 122, 123]. Leveraging the expertise of researchers within the same community, who share similar research interests and experiences, hybrid solutions can offer a tailored approach to generating accurate and descriptive alt text for scientific figures.

The effectiveness of encouraging individuals to contribute to accessibility efforts hinges

on their willingness and intrinsic motivation. Rather than imposing additional burdens, our research focuses on seamlessly integrating accessibility contributions into activities that individuals are intrinsically motivated to perform. Researchers has been explored the concept of "organic" crowdsourcing [124, 125, 126], performing the additional human computation element enhances rather detracting from the main activity, while also resulting in socially useful byproducts. For instance, behaviors such as reading while actively searching for answers and creating annotations for figures can serve as components of a human-in-the-loop AI system to generate vivid descriptions for visually impaired individuals. By embedding accessibility tasks within routine activities, this symbiotic approach enhances the understanding of sighted researchers and facilitates equitable information exchange among all community members including BLV researchers.

Despite the promise of symbiotic information sharing, challenges remain. Designing AI-moderated scenarios that benefit all stakeholders requires dedicated research efforts. As part of this endeavor, we propose Colla, a social reading tool tailored for research papers in HTML format, sourced from the ACM Digital Library. This tool supports figure-centric reading and reflection for sighted researchers while leveraging their interactions to generate interactive, vivid figure descriptions for BLV researchers. By exploring this application scenario, we aim to glean design insights for AI-mediated information communication across diverse stakeholder needs within the same community.

This paper presents findings from a mixed-methods study involving interviews and think-aloud sessions with both sighted (N=7) and BLV researchers (N=7). We discuss the implications of our findings and highlight potential directions for developing a social reading tool that benefits both groups. Our goal is to create a tool that not only enhances accessibility for BLV researchers but also enriches the reading and comprehension experiences for all researchers through collaborative input and advanced AI assistance. The concept was well-received by both communities, showcasing its potential for further investigation and implementation. We discuss lessons and advice as takeaways to share with the community.

In summary, we aim to answer the following research questions:

- What are the challenges in designing a figure-centric, win-win paper reading experience that connects and benefits both sighted and BLV researchers?
- What are the limitations of current AI models, and how can human-in-the-loop systems improve selective text extraction and description generation?

4.2 Related Work

In this section, we looked into prior research on reading tools for general users, assistive reading tools for blind and low vision users, and reviewed AI models for the use case.

4.2.1 Reading Tools for General Users

Reading academic papers is a fundamental part of higher education and research. Digital technologies enable researchers to use reading assistant to help them make annotation and understand the content of the articles[127]. According to prior work [128], paper in office environment are commonly used to print the digital reading material out for active reading which includes tons of annotations to improve content comprehension or to take note for later reference. An earier research done by O'Hara *et al.* [129] provides recommendations for designing digital active reading tasks, emphasizing the importance of annotation and effortless navigation. As one tool to keep reader actively engage with the materials without getting lost, annotation contributes to enhancing reading comprehension [130, 131] . Users can improve their reading experience with SQ3R (Survey, Questions, Reading, Recite, Recall) reading method [132] and reading support activities (Underlining, note-taking, outlining, and networking) [133].

In particular, recent work has highlighted the demand for assistive interactions from reader communities when engaging with new articles [134, 135, 136, 137]. Chen *et al.* [138]

design a human-AI collaborative tool that employs a suite of natural language processing models to provide end-to-end support for reading online news articles. For academic research, critical thinking skill is one of the most significant attributes of a researcher. However, it is not a trivial skill to develop, normally requiring critical guideline [139, 140] or seeking help from experts such as course intructor, senior researchers [141, 142, 143, 144]. CriTrainer [145] is an interactive tool for critical paper reading CPR training. It leverages text summarization techniques to train readers' skills in grasping the paper's main ideas. It further utilizes template-based generated questions to help them learn how to raise critical thoughts. Tan *et al.* [146] presented WiREAD, a web-based collaborative platform that supports peer interactions and provides feedback for both students and teachers to engage in critical paper reading together. WiREAD is shown to enhance the critical reading engagement levels of peers and instructors in their user study. Some work has focused on facilitating discussion of papers by readers, for example by embedding discussion threads within academic papers. The result indicated that such situated discussion can lead to a better understanding of the contents and facilitate learning between peers [146][147]. Other work has focused on enabling researchers to find relevant content, in particular, previous work has created links between visualizations and referencing texts in academic articles [148] and between definitions of terms and their usage in a paper.

4.2.2 Reading Tools for Blind and Low Vision Users

Blind and low vision (BLV) researchers use screen reader to read out the content of paper [149], however, previous researchers have noted that lack of general accessibility has also led to paper abandonment. Wang *et al.* [119] found that academic papers were highly inaccessible for researchers with visual impairments and investigated the challenges and practices of visually-impaired researchers and proposed a system that converts academic papers in an accessible format for the researchers. To consume figures non-visually, alt-text, which is a description, may be embedded in a paper PDF. These descriptions are created

through interpretation by sighted people or conversion of contents into a simplified form [150]. Currently, alt text and image description of academic paper are produced manually and voluntarily by the person who uploaded the image or the authors of the academic paper. Alt-text is often missing, and typically lacks details and structure needed to understand images without viewing them [151]. Automated image description tool has been playing a crucial rule in helping BLV readers to get equitable access to information [152] but auto generation of alt-text for visual components on figure still does not live up to quality needs of researchers.

Park *et al.* [153] designed a reading interface on research papers that enables all readers to contribute to the backend by leaving traces of their reading paths for others to leverage, which benefits both low-vision and sighted users. Zhang*et al.* [154, 155, 156] has been working on supporting accessible slide reading and authoring for blind users. The goal of the ScholarPhi project [157] was to improve the reading of scientific papers by helping readers see the meanings of mathematical symbols, technical terms, and other information directly where they are used within the paper. Lundgard *et al.* introduce a conceptual model for the semantic content conveyed by natural language descriptions of visualization [158], spans four levels of semantic content: enumerating visualization construction properties (*e.g.*, , marks and encodings); reporting statistical concepts and relations (*e.g.*, , extrema and correlations); identifying perceptual and cognitive phenomena (*e.g.*, , complex trends and patterns); and elucidating domain-specific insights (*e.g.*, , social and political context).

4.2.3 LLM-based Techniques in Assisting Research Reading

Large language models (LLMs) are becoming crucial for developing sophisticated agents that can reason, use tools, and adapt to new information across various tasks [159, 160, 161]. As the number and complexity of tasks that can benefit from LLMs grow, a natural strategy for enhancing agent capabilities is to employ cooperative multi-agent systems. Previous research indicates that using multiple agents can foster creative thinking [162], enhance

factual accuracy and reasoning skills [163], and facilitate verification processes [164]. Since GPT-4V start to support multimodal input, it is very natual for human user to "point" something to LLMs [165]. To refer to an arbitrary spatial region of interest, multiple means of interactions including crop box, arrow, box, circle, hand drawing and spatial coordinate can be used as a new prompting methods [166]. However, LLM-based systems still face limitations in retrieving useful information due to context length restrictions [167], which hinders their performance in tasks requiring extensive background information.

LLMs, though powerful, are prone to hallucinations [168, 169, 170] and are restricted to the out-of-dateness of the training data [171, 172]. To address this, Retrieval-Augmented Generation (RAG) [173] has been introduced to bridge the gap between LLMs and real-time data from users. For complex tasks in the context of research reading that require deep and up-to-date knowledge, building systems that integrate LLMs with continuously updated external knowledge repositories is essential. This approach enhances factual accuracy, increases the reliability of generated outputs, and helps reduce hallucinations [168, 169].

4.3 Method

We conducted a mixed-methods user study with 7 BLV participants and 7 sighted participants to *(i)* understand their current practices in reading figures in research papers, and *(ii)* elicit design insights of developing a social reading tool to promote figure accessibility.

4.3.1 Participants

We recruited the BLV participants from an online discussion group email list. We recruited the sighted participants from the authors' institution. We tried to invite participants from different research fields but all read research papers regularly every week. Demographics info of participants is shown in the table 4.1.

Table 4.1: Demographics of BLV and sighted participants in the study. All participants read research papers often and regularly, more than one paper every week and have research experience in their fields over 3 years.

ID	Gender	Age	Vision Status	Research Field
P1	F	39	Totally Blind	Education, Museum
P2	F	27	Totally Blind	Human Rights
P3	M	46	Blind with Some Light Perception	Accessibility
P4	F	33	Low Vision	Education
P5	M	28	Totally Blind	Human Rights
P6	F	38	Totally Blind	Social Psychology
P7	F	34	Totally Blind	Law, Policy
P8	M	23	Sighted	Analog Circuit
P9	M	32	Sighted	Quantum Computation
P10	F	23	Sighted	Medical Imaging
P11	F	27	Sighted	Human-Computer Interaction
P12	F	24	Sighted	Psychology
P13	M	24	Sighted	Statistics
P14	F	22	Sighted	AI for Education

4.3.2 Protocol and Data Analysis

The study consisted of two-stages. In the first stage, semi-structured interviews were conducted to understand participants' current practice and expectations on assistive tools. In the second stage, participants performed think-aloud reading on a paper of their research interests with our demo as a design probe, discussed future improvements on features and attitudes towards social reading for figure accessibility. Deploying a social reading tool for both sighted and blv users in a real-world setting requires significant engineering effort and time. To understand the challenges and opportunities of promoting the accessibility of paper figures through social reading, we decided to first develop a demo with basic functions as a design probe to gain insights and guide the future deployment. The think-aloud study allowed us to gain deeper insights by observing firsthand how participants read figures with assistive tools to further complement the findings from interviews.

All sessions were conducted online, and the total time required for both sessions was between 90 and 120 minutes. With user consent, the study sessions were audio recorded, and the recordings were then transcribed for thematic coding. The study was approved by the ethical review board at the authors' institution. We followed an open coding approach, in which we first independently coded the data and then discussed the codes to reach a consensus. Afterward, we performed affinity diagramming to group the codes into tentative themes and iteratively refined them.

4.3.3 Design Probe

We implemented our design probe as a Chrome Extension, Colla, working on the ACM Digital Library HTML reader view. It activates the Chrome side panel as the interface for assistive functions and uses Node.js as the back-end server. For language model, we use the "gpt-4-vision-preview" model from OpenAI API. Users can navigate between images by clicking a button, which will automatically scroll the paper HTML page to the next figure.

Select the figure will activate the side panel, load the figure and its caption, with assistive functions shown under. The extension can store users' interactions as a JSON file which can also be loaded manually.



Figure 4.1: The demo activates a sidebar in the right side, supporting figure-centric assitive functions including highlights and annotations, questions, and summary. The left side shows the original HTML paper webpage.

4.3.3.1 Highlights and Annotations

Users can highlight the texts related to the figure by left clicking and select then right click to confirm. The highlight of currently selected figure will be shown as *highlight* in the main page while the highlighted text of other figures will be shown as *opaque* to distinguish between highlighted texts of different figures. The highlighted contents are also showed on the right side with the figure to support quick navigation. Users can write down annotations about the figure, also crop a bounding box to select part of the figure as attention focus, with the format shown as follows:

"x":0, "y":0, "width":0, "height":0, "annotation":"Your note here"

4.3.3.2 Questions

Related questions are generated on selected figure based on its visual components and the corresponding context (caption, highlighted texts). The questions are divided into two categories: descriptive questions and general questions. Descriptive questions emphasize the visual attributes in the figure, while general questions focus more on high-comprehension. Users can also customize questions request, ask their own questions, and acquire auto-generated answers as reference.

4.3.3.3 Summary

The demo generates a summary of the knowledge associated with a figure by combining reading behaviors such as highlights, annotations, and question-answering. We introduced Retrieval Augmented Generation (RAG) to integrate reading behaviors from sighted researchers as a knowledge source in the vector database. This approach ensures that updates from researchers are continuously tracked, minimizing misalignment and the risk of misleading data due to the absence or obsolescence of information. For sighted researchers, the prompt focuses on extracting and summarizing highlights and annotations, capturing the researchers' insights and key points related to the figure. For blind researchers, the prompt would also emphasizes explaining the visual components, ensuring that the generated content includes detailed visual descriptions retrieved from the database.

4.3.4 Think-Aloud with the Demo

We asked the sighted participants to read one research paper of their interest using the tool on their laptops by loading the extension package into Chrome. They were instructed to follow the flow of completing figure-centric reading behaviors and to explain their thought processes while reading. The BLV participants also selected a research paper of interest, and we simulated ideal social inputs by reading the paper with detailed figure-centric behaviors.

Because current screen readers do not efficiently support navigating between sidebars, we focused on gathering feedback from the generated contents. BLV participants accessed the paper webpage without the tool, and we conducted the study by inputting their questions and sending generated contents to them.

4.4 Results

For this exploratory study, our results focus on qualitative findings and observations. This study aims to define the aligned and misaligned needs of sighted researchers and blind researchers on paper figures, and the potential functions of AI as moderators in such multi-stakeholders paper reading scenario.

4.4.1 Understanding Sighted Researchers

We need to understand how sighted people leverage figures in their current practice of reading papers. For instance, first skimming through figures and then looking into details they are interested in. And what kind of assistance they think may be helpful for figure reading. And understand what would be the outcome of their natural reading behavior.

4.4.1.1 Figures as Visual Cues in Reading Papers

Participants rely on figures for quick comprehension. After extracting main points or raising questions with a glance at the figure and caption, they tend to look for cross-reference texts for a deeper understanding. To support fully understanding of a figure (*e.g.*, a teaser figure giving an overview of both front-end interface and back-end processing), participants looked into texts in related sections explaining implementation details without direct reference to the figure. In some research fields such as statistics, the figures are mainly informative as data visualizations of showing analysis results, participants mentioned they would jump to

the figure only if they were interested in the results. Meanwhile, in research fields that emphasize visual representation of ideas such as Human-Computer Interaction, papers often contain "summarative" figures that provides overview of sections, participants mentioned they would skim through all figures to grab an overall idea of the whole paper, then decide if they want to read more deeply.

4.4.1.2 Natural Reading Behaviours Motivated by Reflection

Participants have different personal preferences for highlighting texts, taking notes, and annotating figures when reading papers. Rather than absorbing knowledge silently through their mind when reading, the motivation of them for spending extra efforts in making outputs is to help them organize and reflect later. Even if some texts in the paper may provide additional information for them to understand the figure, they may not record it as long as they already get the information. Instead, they judge if such information is important enough for reflection, only if so they would spend efforts in recording these. Without intervention, their notes can be ambiguous, containing only key words but could be easily understood by themselves. Especially, P11 mentioned that she found the figure annotation function very useful, since in her current practice, she couldn't find a reading tool to attach notes to a subpart of figure. But it is also noted that not all participants would like to annotate on figures.

4.4.1.3 Questions as Prompts for Thinking

Participants raised questions on figures based on their expertise and how familiar with the paper topic. Not all questions they have was highly related and really important to the figure content, some are just asked by curiosity, and they will move on as soon as they get the answer; some can be usefully to note down for later reflection. Rather than figuring out the answers by themselves, they hoped to have AI assistance in providing a summarative answer

and point them to related texts in the paper. Participants found the generated questions could help them better understand the paper in different ways. While P9 expressed strong preference on the opened-ended questions that not only restricted in key points from the figure itself, P10, P11, P13, P14 mentioned the visual descriptive questions focuses more on the concrete details and remind them into some insights they would have neglected.

> *According to the generated questions, I start to notice the detail that I may miss without using the reading assistant. I think they (visual components in the figure) may be not that important at first. However, they may be mentioned or become crucial in the following narrative. Being notified their potential significance in advance could help me have a better understanding on the content of the article and save time from navigating back and forth* (P13)

Same as highlighting or taking notes, they are not motivated to manually fill in answer, as long as they grab the answers by themselves; what motivated them is if something is really useful to note down for memorize and future reflections. They feel they may not be able to provide the most accurate answers as well.

4.4.1.4 Summary for Self-Reflection

Participants highlighted the potential of leveraging their reading behaviors to generate summaries that assist with later reflection and recall.

> *Interesting. This summary combines both the content in the paper and my note. I do not know if the content will be still correct if this is integrated with my own idea. Idea is easy to fade away, but this can help me recall info and my thought when I am reading the generated summary since it can remind me of both the content related to this figure and my idea that time.* (P10)

P10 appreciated the ability to merge their notes with the paper's content, indicating

that this combination could aid in recalling both the figure's details and their personal reflections. This integration provides a mechanism for retaining fleeting ideas and contextual information, enhancing their overall understanding and memory of the material. In contrast, P11 expressed a preference for an alternative to the current textual summaries. They found that long textual summaries could obscure their concrete inputs. Instead, P11 suggested generating visual summaries, where her inputs are directly attached to the figure. This method would allow her to easily reflect on the figure and her thoughts upon revisiting it. She emphasized that a visual summary could make it simpler to recall and understand her previous insights without being overwhelmed by lengthy text.

4.4.1.5 Contributing to Accessibility with Social Reading

We didn't disclose to the users that all their reading behaviors may be leveraged to generate descriptions, and as helpful questions-answering to blind peers. They are surprised that they would be able to contribute to figure accessibility in this way. They are concerned about the quality of their inputs. They suggested to allow them to set their reading behaviors as private and public. They would form their language in a more readable way than self notes. Participants also felt excited about how such tool can foster communication in the research community, no matter the experienced level, vision status or other diverse reading abilities. Novice researcher can also get insights from other experienced researchers who are more familiar with the paper topic. Blind researchers can get visual information, have their questions answered by a peer researcher in the same field. For highlighting, annotation, and questions, they would do it in the full paper range, rather than figure-centric (starting from the figure to do such things). The tool can support free-form reading support, then ask the user to select and attach their input with figures to generate a visual summary.

4.4.2 Understanding BLV Researchers

In this section, we explore the unique challenges faced by BLV researchers when interacting with figures in research papers. Through multiple subsections, we discuss their current practices, the spectrum of descriptions required for effective understanding, their attitudes towards AI-generated content, and the benefits of a social reading tool that integrates AI and human inputs.

4.4.2.1 Current Practice with Understanding Figures

All seven participants reported understanding the figure as the most time-consuming part when reading a research paper. They reported in most cases, the papers they read didn't have any embedded alt-text or descriptions for the figures. P3 even mentioned he would directly skip the figures when reading papers since he believes there is nothing readable there. One major gap of experience mentioned by blind researchers is how sighted peers can grasp the gist of the paper by skimming through abstract, title and figures, while they may need twice or third times to get the key points, and hard to verify if their take-aways are correct.

To understand the figures, participants reported the following methods: a) reading the nearby paragraphs above and under the figure to find out more insights; b) they may intentionally search reference about one figure during reading that provide some insights, occasionally found other without reference; c) they leverage AI tools such as chatGPT to provide descriptions and ask questions about the figure (P2, P3, P5, P6); d) they will reach out to sighted peers or helpers to ask specific questions, but challenge is hard to find someone with similar backgrounds.

While P6 mentioned using chatGPT for a special case: accessing tables in paper that are embedded as images that screen readers cannot support navigation the elements in the table. P2, P3, P5 shared similar practice of interacting with chatGPT for multiple-rounds of

adjusted prompting. But if they aquire the information they want or the information they should know still depends on if they are able to ask the right questions. We attached two exampler concrete interaction process provided by P2 and P5 in the Appendix.

4.4.2.2 A Spectrum of Descriptions for a Figure

> *What are we talking about when we talk about 'description'? it can be a spectrum depends on multiple dimensions rather than a fixed format.* (P3)

When discussing their definitions of a good description for a research paper figure, participants emphasized the need for customizable descriptions rather than a one-size-fits-all standard. They pointed out that various factors—such as different figure categories, reading goals, expertise in the paper's topic, and whether visual impairments were born or acquired—can influence how descriptions should be tailored. Based on participants' input, we summarized this into a three-level framework:

(i) First-level (Short Alt-Text): This provides a brief overview of the figure, including its category (photograph, illustration, bar chart, flowchart, mixed) and its significance to the paper's theme (informative, summative, decorative). At this level, users can quickly decide whether to spend more time examining the figure.

(ii) Second-level (Contextual Description): This level offers more contextual information, visual details, and inferences, helping to fill the visual gap for the user. Specially including the intention of author hope to convey with the figure, and concepts that highlighted by the authors with special visual designs.

(iii) Third-level (Interactive Description): This level meets customized needs through interactions, including question-answering and navigating through the figure. It allows users to explore details independently and verify information, not only by understanding the figure itself but also by placing it within the broader context of the paper, such as asking about a term mentioned in the figure.

If a figure includes multiple subfigures, the description of the subfigures should follow the same framework.

4.4.2.3 Attitudes towards AI Generations

Overall, BLV participants showed positive attitudes towards using AI as an always-available assistive tool for understanding paper figures. However, they also expressed a desire not to over-rely on AI and be isolated from sighted peers by AI tools, emphasizing the importance of human interaction and collaboration. Based on interviews and observations, several conditions were identified where AI struggled to perform effectively:

(i) Complex Figures: Research paper figures often contain multiple components that AI may not process well. For instance, data visualizations and flowcharts can pose significant challenges. AI frequently makes errors when describing continuously changing trends in data if it relies solely on the figure. However, it can correct these errors when supplemented with related textual descriptions from the paper.

(ii) Scanned PDFs and Low Contrast: Figures in scanned PDFs of older papers can be unclear, making it difficult for AI to classify them accurately. Similarly, figures with low contrast color designs pose challenges for AI interpretation.

(iii) Layout Segmentation: AI tools like ChatGPT currently struggle with precise segmentation of figure layouts, raising questions about their ability to generate navigation lists for flowcharts effectively. Additionally, AI sometimes generates content not present in the paper, based on its own understanding, which can be misleading.

For important conclusions, participants typically repeated prompts to the AI and compared the inconsistencies in the results to make their own judgments. They expressed a need for efficient access to human verification for AI-generated content, particularly from peer researchers in the same field. They suggested that if AI is making conclusions, additional verification methods should be available. These methods could include providing a confi-

dence score for conclusions, explaining and pointing to the original data supporting these conclusions, and facilitating navigation through the data.

P3 highlighted the ongoing debate in the literature regarding the subjective versus objective nature of descriptions. Describing a trend that the author aims to show can be subjective, whereas describing an element's color is entirely objective. However, purely objective descriptions can render scientific paper figures nonsensical. P5 expressed a preference to avoid subjective opinions in descriptions, especially when describing people. They were, however, open to showing trends in data visualization, provided there is access to the original data or other ways to verify the results.

4.4.2.4 Benefits of a social reading tool

After the study conductor explained how the tool could be supported by inputs from other researchers reading the same paper, all participants expressed strong enthusiasm for such a social reading tool for research papers. Participants found the ability to see related text highlights from other researchers particularly helpful. This feature was compared to Kindle's highlighting and note-sharing functions. By checking sighted researchers' highlights and question-answers, participants could gain insights into the paper's core ideas and receive cues on navigating the paper's structure.

The hybrid solution of AI and crowd-sourced inputs was highlighted as a beneficial model. For example, the application Be My Eyes and its new function Be My AI were cited. Be My Eyes is a mobile app that helps blind and visually impaired people recognize objects and manage everyday situations with the assistance of sighted volunteers via live chat. Be My AI, powered by GPT-4, provides detailed descriptions of pictures and allows users to chat and ask further questions. If Be My AI cannot answer a question or if users need additional assistance, they can still connect with human volunteers.

P3 emphasized that such a tool can be inclusive and beneficial for researchers with

different abilities, fostering a new culture in the research community:

> *"We should forget about specialized assistive tools and move towards mainstream technologies that can be adapted to the needs of various people...There could be contributors, for example, when an author or a researcher reads and contributes maybe one minute of their time answering questions about a figure. The AI can then enhance the descriptions it has already generated over time, making the content richer and richer."* (P3)

P2 expressed anticipation for how this tool could foster connections and collaboration between sighted and blind researchers, making education and research more accessible in resource-limited areas. P7 also believed this tool would benefit people with various impairments, including ADHD, reading difficulties, and color blindness.

4.5 Discussion

Our study revealed challenges and opportunities of promoting accessibility of paper figures by social reading which can benefit both sighted and BLV researchers in the same community. In this section, based on the results above, we define aligned and misaligned needs from both group, discuss the collaborating role of AI and crowdsourcing, and future work on improving development and deploying the tool in the real world setting.

4.5.1 Mapping Needs of BLV and Sighted Researchers

BLV and sighted researchers share overlapping needs when it comes to understanding figures in research papers. The ultimate goal is to place the figure in the context of the paper to fully comprehend its content. We discuss the specific challenges BLV researchers meet under different conditions as follows.

First, even when the information conveyed by a figure is fully covered by the text in

the paper, BLV researchers still face unequal access because they need significantly more time to locate and acquire the same amount of information as their sighted peers. Second, figures can convey information not explicitly covered in the text. For example, a flowchart illustrating the step-by-step logic of data processing can provide intuitive understanding, which the author might not describe in detail within the paper. Third, figures often use color contrast and other visual designs to emphasize or convey certain information quickly, which sighted readers can grasp instantly. However, blind individuals might need to spend more time reading the text to infer the same information.

It is also worth further exploration on how figure-centric reading may benefit sighted researchers. In our study, sighted participants mentioned that visual questions can sometimes prompt deeper thinking by focusing on specific visual concepts in the figure. However, these questions can also be unhelpful if they focus too much on basic visual attributes. Determining how to frame effective visual questions and evaluate their helpfulness requires additional experimentation.

Without clear motivation for sharing and fostering discussion, natural reading behaviors may not yield useful inputs for others. Participants noted that if visual information is easily understood at a glance, they are often not motivated to document it, instead keeping it in mind. For basic visual details, AI can serve as a preliminary solution, generating descriptions that sighted researchers can lightly verify. Furthermore, sighted researchers can be encouraged to contribute more valuable information to blind researchers, such as by adding prompt questions to ask researchers to annotate the most important parts of the figure or tagging their highlights or notes to the figure. There also exists potential of leveraging AI to reframe their free-form language as qualified and clear inputs to the knowledge database based on the context of the paper.

4.5.2 Building Collaborative Intelligence of AI and Crowd

The information related to a figure may be scattered throughout the paper, not just in sentences explicitly citing the figure (*e.g.*, as shown in Figure X). For instance, to fully understand a summarative figure showcasing a system, researchers may need to refer to implementation details found elsewhere later in the text. Currently, it is still challenging for AI to infer all such ambiguous information from the lengthy context of a paper based solely on one figure. With human researchers' highlights serving as additional manual "labelled data", AI can generate a more comprehensive summary for understanding the figure.

More importantly, the collaborative system should consider the limitations of current visual language models, such as GPT-4 with vision behind ChatGPT, when interpreting figures in research papers. These limitations include difficulty in reading small text within figures (a common occurrence in research papers), challenges with visual elements that have varying colors and styles (often seen in data visualizations, summary teasers, or conceptual figures), and struggles with spatial reasoning, such as precise spatial localization. This spatial reasoning is crucial for supporting spatial and subcomponent mind mapping, which is essential for BLV researchers to fully understand the figure. The model may also generate incorrect descriptions or captions in certain scenarios, presenting them as accurate, which makes it challenging for BLV researchers to discern the truth. In multiple rounds of prompting, the model may start answering questions based on its own training data, which can be outside the scope of the paper, thus providing false information, as demonstrated in the examples in the appendix. These limitations can be addressed by designing a social reading interface that integrates AI with inputs and monitoring from human researchers.

While models are often trained on clean, formatted, and well-labeled data, there are both challenges and opportunities in exploring how the massive and chaotic crowd reading behaviors can be parsed and organized to form a constantly developing knowledge base for a figure. This evolving knowledge can be leveraged by AI to ensure stable and consistent

outputs based on a consensus, while carefully avoiding the deepening of biases and fostering constructive discussion and communication. The absence of a benchmark dataset of good descriptions for different kinds of figures also presents a challenge. We need to embed the values and standards of figure accessibility into the prompts used by models to generate descriptions. Furthermore, evaluating performance on a large scale and in a qualitative way, rather than relying solely on quantitative subjective ratings from BLV researchers on a small dataset, is crucial. In an ecosystem with multiple stakeholders (AI, researchers with different abilities), it is important to co-design AI systems that address both shared and unique needs. Such systems should facilitate each users' contribution to the community and provide benefits to other stakeholders.

4.5.3 Crafting a Unified Tool for All Users with Different Abilities in One Community

A single large language model's (LLM) extensive range of capabilities can be effectively combined in a modular way through conversations between agents with different configurations, representing various stakeholders. We should aim to develop a multi-agent conversation framework specifically designed for a social reading tool, with a general abstraction and an effective implementation that meets the diverse needs of these stakeholders. Each agent must be tailored to handle specific tasks relevant to different user groups. For instance, an agent for blind researchers may emphasis the visual cues they are not aware of, while lower the importance ranking and avoid repeating such info which can be accessed by sighted researchers at a glance.

We should also consider how different researcher roles and subgroups within the community may benefit from such a social reading tool. Novice sighted researchers unfamiliar with a topic may also struggle to understand figures, particularly when they represent novel contributions. In such cases, these researchers face similar challenges as their BLV peers in semantically understanding the paper. As they read, they can gain more insights, and

annotations or highlights from other researchers can provide efficient navigation and helpful insights for both BLV and sighted researchers. For senior researchers, the tool can offer a platform for sharing expertise and mentoring others. It can also facilitate collaborative research by allowing them to annotate and highlight figures, providing valuable context and insights for their peers. Authors can benefit by learning about the diverse understandings of different audiences. The tool can help them design accessible figures, write clear captions and descriptions, and provide adequate references to figures within the paper text. It also allows authors to add accessibility features to previously published papers that lack alt-text when resubmitted. BLV reviewers also play a critical role in the peer review process of publications, and this tool can greatly enhance their ability to assess research papers effectively. Additionally, this tool can be adapted to support group collaborations and peer-to-peer paired reading, enhancing the research experience for all users with different abilities. It can foster a more inclusive and interactive research community, enabling better communication and understanding across different levels of expertise and abilities.

4.5.4 Limitation and Future Work

We took the first step to understanding how BLV people perform through a mixed-method approach of a think-aloud task session and an interview. Our work provides initial insights into promoting figure accessibility with social reading. More efforts are needed to make papar figures more accessible for BLV people so that they could tap into the power of research reading along with their sighted peers. We highlight a few limitations of our current work and future research directions. While we provide largely empirical findings and design insights based on the user study, the study have several limitations. The study had small numbers of participants, largely due to the scarcity of low-vision researchers and resulting recruiting difficulties, which precluded quantitative inferences. We also only deployed our system with selected papers by the research team and participants. This constraint limits the opportunity for us to see how users might actually use the system in the wild with papers

of their own choosing. Future work should examine how the system can be utilized in the longer term involving more participants. Further studies with larger pools of participants and more design probes are needed to more fully understand user experiences.

CHAPTER 5

Summary

This dissertation focuses on engaging internet users in making visual media accessible at three progressive levels: *(i)* leveraging information that people unintentionally and randomly contribute to accessibility; *(ii)* nudging people to intentionally and proactively contribute to accessibility, with additional effort; *(iii)* creating social information-seeking experiences where people naturally contribute to accessibility, aligned with their original goals. Three representative application scenarios are presented: online shopping, online video watching and commenting, and online research paper reading, each corresponding to one of the three levels. By grounding these engagement goals in scenarios that urgently need improved accessibility, this work elicits concrete implementation strategies.

Chapter 2 presents Revamp, an interactive information retrieval system that integrates with Amazon and leverages customer reviews to support accessible information seeking of BLV users. To leverage information that people unintentionally and randomly contribute to accessibility, my work offers novel insights into accessible information seeking by extracting visual information from existing human-authored texts in the source where the visual media is found.

Chapter 3 presents CoSight, a Chrome extension designed to augment YouTube video pages with interactive nudging features to support accessible comment authoring by sighted audiences. To nudge people to intentionally and proactively contribute to accessibility, my work explores the design of lightweight interventions by nudging sighted people to contribute to accessibility by using visual descriptive language in online posts.

Chapter 4 introduces Colla, a social reading tool for online research papers that connects and benefits both sighted and BLV researchers. To create social information-seeking experiences where people naturally contribute to accessibility, my work embeds the value of accessibility into social information foraging by designing a unified tool that benefits both sighted and BLV individuals within the same community.

5.1 Limitations and Future Work

This dissertation explores only one level of user engagement in each representative application scenario. However, in real-world deployment, there might be a mix and flexible adaptation of none, one, two, or all three levels of engagement based on various platforms, media types, contexts, and user archetypes. These archetypes [1] consider users' awareness (what they know), concerns (how they care about it), and expertise (their skills or knowledge to deal with it). Future efforts are required to develop implementation strategies for handling diverse situations.

From Level I to Level III, this dissertation transitions from focusing on a single targeted user group (BLV users) to envisioning a future of mixed-ability research [174, 175, 176]. In this future, multiple stakeholders with different abilities and needs are collaboratively and equally involved. The future design of mainstream platforms and techniques should take into account their use in different contexts where people with varying abilities may leverage them to fulfill tasks otherwise impossible or difficult for them [177].

This dissertation also lays out a vision for interactive computing systems that combine machine intelligence with crowd intelligence to promote social good. It applies the ideas of nudging and organic crowdsourcing proposed by prior research into the largely unexplored area of promoting accessibility. It moves beyond crowd-powered systems [178] to user-advocated systems, engaging platform users themselves to contribute to the platform's

[1] proposed by Melody Ivory in UIST 2021 Keynote Talk: More Human HCI.

accessibility. While this dissertation focuses on user engagement, it does not deeply explore important ethical issues that must be carefully examined in real-world settings, such as manipulation threats, abuse, and potential bias.

As machine intelligence rapidly advances, some specific gaps that machines currently cannot automate may become less relevant. However, the long-term value of tightly integrating user advocacy and internet accessibility remains significant. Rather than being isolated by AI tools, BLV people and sighted people can work together to co-create accessible households [114], workspaces [115], and online communities.

The deployment and small numbers of participants numbers in some studies, though adequate for initial exploration, might not capture the full spectrum of user interactions and experiences. Future work should examine how the systems can be utilized in the longer term involving more participants in real-world settings.

5.2 Closing Remark: Towards a Unified and Inclusive Internet

The Internet has radically transformed our lives for the better, creating virtually infinite opportunities for its users. Yet, what makes the Internet great also makes it vulnerable. When a solution is meant to serve any member of society and then doesn't, these mismatched designs contribute to the societal invisibility of excluded groups [179]. Inclusion is not only for those who are currently excluded but for everyone. We are accountable when someone is left out, just as we are when it happens to ourselves. It is up to all of us to create a unified and inclusive Internet that everyone in the world can access and benefit from.

APPENDIX A

Appendix to Chapter 2

A.1 Demographics of study participants

20 participants were involved in the studies.

Table A.1: The demographic information of participants.

ID	Age/Gender	Visually impairment	Device
P1	24/F	Blind (total), acquired	smartphone
P2	24/F	Blind (some light perception), born	smartphone
P3	22/F	Blind (some light perception), born	smartphone, laptop
P4	25/M	Blind (total), born	smartphone
P5	24/M	Blind (some light perception), born	smartphone, laptop
P6	24/M	Low Vision (narrowed field), born	smartphone, laptop
P7	27/M	Low Vision (narrowed field), born	smartphone
P8	32/M	Blind (total), born	smartphone, laptop
P9	28/M	Blind (total), born	smartphone, laptop
P10	22/M	Blind (total), acquired	smartphone, laptop
P11	27/M	Blind (total), born	smartphone,laptop
P12	23/F	Low Vision (narrowed field), acquired	smartphone, laptop
P13	34/M	Blind (total), born	smartphone, laptop
P14	35/M	Blind (some light perception), acquired	laptop
P15	33/M	Blind (total), born	smartphone, laptop
P16	22/F	Blind (some light perception), born	smartphone, laptop
P17	27/M	Blind (some light perception), born	smartphone, laptop
P18	31/M	Blind (total), born	smartphone, laptop
P19	26/F	Blind (some light perception), born	smartphone, laptop
P20	44/F	Blind (total), acquired	smartphone, laptop

P1 - P10: Chinese Users P11 - P20: North America Users

APPENDIX B

Appendix to Chapter 3

B.1 Instructions on writing accessible comments with CoSight

B.1.1 Why your comments may help blind people

Describing visual information can help blind and visually impaired people who cannot see the video adequately. Professionals add descriptions to the original audio tracks for films and some videos. It is not realistic to have every online video described by professionals or content creators, however, every ordinary audience like you has the power of contributing to the accessibility of the video you are watching, by making your comment more accessible by clarifying the visual, or writing a descriptive comment, especially for an inaccessible segment! When blind people are listening to the videos, we make a beep sound if helpful comments are available for a segment, then they can choose to pause the video, use screen readers to navigate, and read out the selected comments.

B.1.2 Make your comment more accessible by clarifying the visual

Add timestamp if you are talking about specific segments of the video; Avoid using ambiguous pronouns, such as it, that, or this, instead, use the full names or concepts you refer to. If you share feelings on visual content not directly covered in the audio track, describe the visual content. Here is an example of modifying a comment:

- **Original**: I like it when he walked in. The scene was so cool!
- **More accessible**: 5:26 I like it when the vlogger's friend walked in. The scene of all the lights turning on behind him was so cool!

Here are more examples of comments with clarified visuals:

- 1:29 WAIT A MINUTE! Taste perception is not localized like that in the brain. Different tastes are not processed in different parts of the cerebral cortex as the illustration showed.
- 1:42 thanks, now my legs are going to be itchy for the rest of the video because of the little animation of the spider moving on the skin

B.1.3 Write a descriptive comment, especially for an inaccessible segment

Visual elements that are important to understand what the video is communicating:

- Objects (e.g. ingredients needed for a recipe);
- Actions (e.g. a step in a tutorial video);
- Scenes (e.g. a person studying);
- Animations (e.g., an arrow showing the direction of force);
- Texts (e.g., additional info not mentioned in the audio).

Don't need to describe every detail: Apparent from the audio. (e.g. host speaking to camera); Already mentioned/described in the audio.

The system may not identify every segment that deserves a description, feel free to add comments to where you think is meaningful!

Here are some examples of descriptive comments:

- 0:28 I learned a little trick here: reverse the pan and use the back of the pan to form the shape of foil
- 0:45 She showed a white pan and a black pan. If you use a black pan, remember to reduce the temperature of your oven by 25 degrees
- 1:39 It seems that she used at least three eggs

B.2 Explaining inaccessible segments and Instructions on What to Describe

The reasons of inaccessibility followed the definitions from prior work [11].

- **R1 Contain no-speech segment**: clarify the visual contents
- **R2 Not informative language**: clarify the concrete details of visual contents
- **R3 Frequent shot changes**: clarify the changing scenes
- **R4 Numerous visual entities**: clarify any important visual entities
- **R5 Specific visual details not covered by speech**: clarify any missing visual details
- **R6 On-screen text not in speech**: clarify the important on-screen text
- **R7 Reference words without explanations**: clarify it/this/that words in the speech

B.3 Comments Collected by CoSight in Study with Sighted Audience

Table B.1: The accessible comments with timestamps and descriptive visual details written by participants.

Video Title	**Accessible Comments Written with CoSight**
The Nervous System, Part 1: Crash Course Anatomy & Physiology #8	8:60 neurons are shown going down the nerves in the image of a dark body from the top to the leg
How to Bathe your Cat that Hates Water (6 Step Tutorial) — The Cat Butler	0:26 a man pours water onto a wet cat inside the bath tub from a pitcher
Fox Village in Japan: The Fluffiest Place on Earth! (Miyagi Zao, Shiroishi)	0:17 There's about 30 rabbits surrounding that woman! Too cute! They look so fluffy and soft and clean.
How sugar affects the brain - Nicole Avena	At 1:12 they are showing a drawing of a tongue and highlighting different parts to show where on your tongue different tastes are felt.
Tigers 101 — National Geographic	0:56 The introductory scenes are of select shots of tigers in the wild, beautiful creatures, and one even looks at the camera.
How to Plant Potatoes! Garden Answer	0:29 She is sitting on the ground by a raised planter. Beside her is a tray of potatoes, and she is touching the dark, loamy soil with her bare hand. Behind her is a pitchfork and the bag the organic gardening soil came from.
UW Campus Tour: Hear It from a Husky	1:16 The red square is such a large open area between multiple buildings, layed with burgandy colored bricks. Very stunning.
Experiment! How Does An Owl Fly So Silently? — Super Powered Owls — BBC	2:05 The owl appears to have red eyes in a zoomed view. The red eyes and white fur give me the impression this is an albino variant.

APPENDIX C

Appendix to Chapter 4

C.1 Example 1: Blind Researcher Interacting with chatGPT

P2 cropped and input the figure 1 from paper ComputableViz [180].

C.1.1 Round 1

Ask

please detailedly describe the uploaded figure to a visually impaired individual. Please explain the potential inplication or meaning of the figure.

Answer

The figure you provided is a conceptual diagram from a scientific or academic publication about "ComputableViz," a library for processing and analyzing visualizations using mathematical operators. This figure represents a design space of visualization operations organized along two dimensions: what can be operated on (operation targets) and how these operations are carried out (operation types).

In this figure: 1. **Operation Targets** are divided into two categories: - **Data-related**: Focused on the data aspects of the visualizations. - **Style-related**: Concerned with the aesthetic and structural components of the visualizations.

2. **Operation Types** include: - **Basic operations**: These are foundational operations such as union, intersection, and difference that can be applied between two visual-

izations. - **Advanced operations**: These involve more complex manipulations and are applicable to scenarios involving more than two visualizations, like sorting or clustering.

The figure likely uses a variety of symbols and colors to categorize and differentiate these aspects visually. For example, different shapes or colors might indicate different categories or types of operations, and arrows or lines might depict relationships or processes.

Implications or Meaning: This diagram serves as a framework for understanding how various operations can be applied to visualizations in a systematic way. It illustrates a structured approach to managing and manipulating data visualizations beyond simple creation or display, aiming to enhance capabilities like automated analysis, integration of multiple data sources, and more dynamic interactions with data. This framework could significantly aid researchers and developers in creating more robust tools for data analysis, thereby enhancing the accessibility and usability of data visualization technology in various fields.

Comment

Not quite satisfied. This is because the description is confused. I am not very sure the structure of the figure. I can only understand vague purpose of why using this figure in the article.

C.1.2 Round 2

Ask

Please detailedly explain the figure structure to me.

Answer

The figure from your document represents a structured framework for understanding and implementing operations on visualizations, categorized into operation targets and operation types, which are foundational concepts in the "ComputableViz" project. Here's a breakdown of the figure's structure based on the detailed content provided:

1. **Operation Targets**: These are the elements within visualizations that operations can be applied to. They are divided into: - **Data-related**: This includes components of the visualization that deal directly with the data itself, such as datasets and transformations applied to the data. - **Style-related**: This involves aspects that affect the appearance and presentation of the visualization, such as markers, encodings, and other aesthetic parameters.

2. **Operation Types**: These define how operations are applied to the targets, categorized into: - **Basic Operations**: These are simple operations applied between two visualizations and include: - **Union**: Combining elements from two or more visualizations. - **Intersection**: Identifying common elements between visualizations. - **Difference**: Determining the elements present in one visualization and absent in another. - **Advanced Operations**: These involve complex processes that manage multiple visualizations and can include operations like sorting and clustering, which help organize and analyze larger sets of data or visual elements.

Each operation target and type is likely visualized using distinct symbols or icons in the figure, possibly with connectors that illustrate potential applications or flows between different operations, showing how one can transition from basic operations to more complex analyses.

This structured approach aids in systematizing the manipulation and processing of visualizations, making it easier to develop tools and methods for complex data analysis tasks. It also highlights the versatility of the ComputableViz library in handling different aspects of visualization beyond mere data representation, facilitating advanced functionalities like automated analysis and interactive data exploration.

Comment

Satisfied. I have a good understanding of the purpose and usage of this figure. The only disadvantage is that this answer didn't provide the objective description to the figure itself, such as its collor or layout.

C.1.3 Round 3

Ask

please detailedly describe each element of the figure, including their inner location, size and contrast. If there are some color changes, please describe and explain their potential inplication.

Answer

Here's a detailed description of the figure from your document, focusing on the elements within the figure, their locations, sizes, contrast, and any color changes:

Figure Description The figure is structured around a central grid that visually categorizes operations applicable to visualizations. Each category is defined by two axes: one for "Operation Targets" (the vertical axis) and one for "Operation Types" (the horizontal axis).

Operation Targets - **Data-related**: Located on the upper half of the grid. This category is highlighted in a lighter shade, possibly light blue or gray, to differentiate it from the style-related category. This visual differentiation implies a conceptual separation between operations affecting the data itself and those affecting its presentation. - **Style-related**: Positioned on the lower half of the grid. It is likely shaded differently, perhaps in a darker shade like dark blue or gray, indicating a focus on visual aesthetics and formatting.

Operation Types - **Basic Operations**: These are shown on the left side of the grid, spanning from top to bottom. This section might use a uniform color to signify basic operations that apply both to data and style, emphasizing their foundational role in visualization manipulation. - **Advanced Operations**: Located on the right half of the grid. This section could be visually distinct from basic operations, possibly using a gradient or a slightly altered hue to suggest a higher level of complexity or a different set of functionalities.

Visual Elements and Implications - **Grid Divisions**: The clear division into four quadrants helps users quickly understand the classification of operations. The use of contrasting shades or colors between different categories enhances readability and focus. - **Text and

Labels**: Labels are crucial for identifying the types of operations and their targets. The text is likely concise and strategically placed to maintain the structure's clarity without overcrowding the visual space. - **Color Usage**: If there are color changes, they are likely used to denote different types of operations or to distinguish between data and style targets. The choice of color can guide the viewer's interpretation, with cooler tones perhaps used for data-related elements to symbolize logic and analytics, and warmer tones for style-related elements to convey creativity and design focus.

Implications The structured arrangement and the use of color in this figure are not just aesthetic choices but serve to facilitate rapid comprehension of complex information. The distinction between operation types and targets through visual means (like color and layout) helps in mentally organizing the applicable operations, which can be crucial for users like data scientists or developers who need to make quick decisions about data visualization strategies.

This detailed description aims to provide you with a thorough understanding of the figure's layout and design principles, emphasizing how visual elements are employed to enhance interpretability and user engagement.

Comment

If the answer were correct, I am very happy with the description. I now got the main points of the figure and can construct the figure in my mind.

C.2 Example 2: Blind Researcher Interacting with chatGPT

P5 cropped and input the figure 2 from paper Revamp [10].

Input

I'm reading a academic paper. There is a figure in the paper and I'm blind. Please describe the figure for me as detailedly as possible.

chatGPT

Certainly, I'll describe the figure for you. The image you've provided appears to be a flowchart with various rectangular blocks connected by arrows, indicating the flow of processes or information. At the top, there are two separate starting points labeled "Voice" and "Text," suggesting that there are two types of inputs to the system or process being described. From both the "Voice" and "Text" blocks, arrows point down to a block labeled "Visual Questions w/ Basic Attributes," indicating that both types of input lead to the generation or consideration of visual questions that have basic attributes. There is another block at the same level as "Visual Questions w/ Basic Attributes," but it's not directly connected to the Voice or Text blocks. It's labeled "Other Questions." This suggests that there are additional questions considered that do not come from the basic voice or text inputs. The "Visual Questions w/ Basic Attributes" block and "Other Questions" block both point to a central block labeled "Special Rules." This could mean that special rules are applied to both visual questions with basic attributes and other types of questions. From the "Special Rules" block, there is an arrow pointing to another central block labeled "Pythia Visual QA API," indicating that the special rules are applied within or lead to the use of the Pythia Visual QA API. To the right of the "Pythia Visual QA API," there is a process that seems to split into two paths. One arrow leads to a block labeled "Reviews Ranking," and another arrow from the same point leads to a block labeled "Summary." This suggests that the output of the Pythia Visual QA API can be used for ranking reviews and for creating summaries. The "Reviews Ranking" block points to a block labeled "Positive/Negative Reviews," indicating that reviews are possibly categorized into positive or negative as part of the ranking process. Finally, the "Positive/Negative Reviews" block and the "Summary" block both point to the last block on the right, labeled "Answers." This seems to be the end output of the flowchart, suggesting that the process results in some form of answers based on the inputs, rules, and analyses conducted throughout the steps. The flowchart is quite linear, with a clear progression from inputs to outputs, and appears to be related to processing and

analyzing questions and reviews, possibly in the context of a question-answering system that handles both textual and voice inputs.

Input

It's difficult for me to understand a flowchart due to my visual impairment. Please explain it with as few visual elements as possible.

chatGPT

Absolutely, let me simplify the explanation: The figure represents a sequence of steps or a process where two types of information, "Voice" and "Text," start the sequence. Both types converge into a stage where they are treated as "Visual Questions with Basic Attributes." Simultaneously, there's a separate strand for "Other Questions" that bypasses the voice and text inputs. Both strands—the visual questions and the other questions—then go through "Special Rules." After passing through these rules, the process enters a central system or tool named "Pythia Visual QA API." From this central point, the process diverges into two separate outcomes: 1. The first outcome ranks "Reviews," classifying them into "Positive/Negative Reviews." 2. The second outcome generates a "Summary" of the information. Finally, both outcomes—ranked reviews and the summary—feed into the final result, which is termed "Answers." Imagine this as a recipe that starts with two ingredients, combines them, applies a method, and then splits the result into two dishes, which are finally served together.

Input

What does "basic attributes" mean?

ChatGPT

"Basic attributes" in the context of a process like the one in the flowchart you described typically refer to fundamental or essential characteristics that define the information being processed. When dealing with visual questions, these attributes might include: 1. Content: What is the subject or main focus of the question? 2. Type: What kind of question is it?

(e.g., is it seeking a definition, an explanation, a description, etc.) 3. Purpose: Why is the question being asked? (e.g., to gain information, to clarify something, etc.) 4. Structure: How is the question formulated? (e.g., what is the grammatical structure or complexity of the question?) 5. Keywords: Are there specific terms or phrases that are crucial to understanding the question? In a Visual Question Answering (VQA) system, basic attributes help the system to understand and categorize questions before applying more complex rules or processing steps. These attributes serve as the starting point for how the system will interpret the question and decide on the kind of answer that should be generated.

Input Really? Is it possible that the phrase refers to color, shape, etc.?

chatGPT Yes, you're correct. In the context of a Visual Question Answering (VQA) system like the one mentioned in the flowchart, "basic attributes" would indeed typically refer to fundamental visual features such as: 1. Color: The hues or shades present in an image. 2. Shape: The geometric form of objects within the image. 3. Size: The scale or dimensions of elements in the image. 4. Position: The location of objects relative to one another. 5. Texture: The surface quality or visual patterns on objects. 6. Count: The number of certain objects or elements. These attributes are often the primary descriptors used to analyze and understand visual content, and they would be the kind of basic information that a visual processing system would initially extract from an image to answer related questions. ChatGPT can make mistakes. Consider checking important information.

REFERENCES

[1] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, and Tom Yeh. Vizwiz: Nearly real-time answers to visual questions. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)*, W4A '10, New York, NY, USA, 2010. Association for Computing Machinery.

[2] Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P. Bigham. Visual challenges in the everyday lives of blind people. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2117–2126, New York, NY, USA, 2013. ACM.

[3] Erin Brady, Meredith Ringel Morris, and Jeffrey P. Bigham. Gauging receptiveness to social microvolunteering. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 1055–1064, New York, NY, USA, 2015. Association for Computing Machinery.

[4] Amy Pavel, Gabriel Reyes, and Jeffrey P Bigham. Rescribe: Authoring and automatically editing audio descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 747–759, New York, NY, USA, 2020. Association for Computing Machinery.

[5] Jeffrey P. Bigham, Richard E. Ladner, and Yevgen Borodin. The design of human-powered access technology. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '11, page 3–10, New York, NY, USA, 2011. Association for Computing Machinery.

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

[7] Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M. Kitani, and Jeffrey P. Bigham. 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, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery.

[8] Abigale Stangl, Nitin Verma, Kenneth R. Fleischmann, Meredith Ringel Morris, and Danna Gurari. Going beyond one-size-fits-all image descriptions to satisfy the information wants of people who are blind or have low vision. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '21, New York, NY, USA, 2021. Association for Computing Machinery.

[9] Richard H Thaler and Cass R Sunstein. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin, 2009.

[10] Ruolin Wang, Zixuan Chen, Mingrui Ray Zhang, Zhaoheng Li, Zhixiu Liu, Zihan Dang, Chun Yu, and Xiang 'Anthony' Chen. Revamp: Enhancing accessible information seeking experience of online shopping for blind or low vision users. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[11] Xingyu Liu, Patrick Carrington, Xiang 'Anthony' Chen, and Amy Pavel. What makes videos accessible to blind and visually impaired people? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[12] Xingyu "Bruce" Liu, Ruolin Wang, Dingzeyu Li, Xiang Anthony Chen, and Amy Pavel. Crossa11y: Identifying video accessibility issues via cross-modal grounding. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA, 2022. Association for Computing Machinery.

[13] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. "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, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery.

[14] Abigale J. Stangl, Esha Kothari, Suyog D. Jain, Tom Yeh, Kristen Grauman, and Danna Gurari. Browsewithme: An online clothes shopping assistant for people with visual impairments. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '18, pages 107–118, New York, NY, USA, 2018. ACM.

[15] Guanhong Liu, Xianghua Ding, Chun Yu, Lan Gao, Xingyu Chi, and Yuanchun Shi. "i bought this for me to look more ordinary": A study of blind people doing online shopping. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 372:1–372:11, New York, NY, USA, 2019. ACM.

[16] Kristen Shinohara and Jacob O. Wobbrock. In the shadow of misperception: Assistive technology use and social interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 705–714, New York, NY, USA, 2011. ACM.

[17] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. Vizwiz: Nearly real-time answers to visual questions. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 333–342, New York, NY, USA, 2010. ACM.

[18] Mauro Avila, Katrin Wolf, Anke Brock, and Niels Henze. Remote assistance for blind users in daily life: A survey about be my eyes. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '16, pages 85:1–85:2, New York, NY, USA, 2016. ACM.

[19] Meredith Ringel Morris, Kori Inkpen, and Gina Venolia. Remote shopping advice: Enhancing in-store shopping with social technologies. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 662–673, New York, NY, USA, 2014. ACM.

[20] Chien Wen Yuan, Benjamin V. Hanrahan, Sooyeon Lee, Mary Beth Rosson, and John M. Carroll. I didn't know that you knew i knew: Collaborative shopping practices between people with visual impairment and people with vision. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW):118:1–118:18, December 2017.

[21] André Rodrigues, Kyle Montague, Hugo Nicolau, João Guerreiro, and Tiago Guerreiro. In-context q&a to support blind people using smartphones. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '17, pages 32–36, New York, NY, USA, 2017. ACM.

[22] Michele A. Burton, Erin Brady, Robin Brewer, Callie Neylan, Jeffrey P. Bigham, and Amy Hurst. Crowdsourcing subjective fashion advice using vizwiz: Challenges and opportunities. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '12, page 135–142, New York, NY, USA, 2012. Association for Computing Machinery.

[23] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, page 15–29, Berlin, Heidelberg, 2010. Springer-Verlag.

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

[25] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018.

[26] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 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*, pages 1180–1192. ACM, 2017.

[27] Christina Low, Emma McCamey, Cole Gleason, Patrick Carrington, Jeffrey P. Bigham, and Amy Pavel. Twitter a11y: A browser extension to describe images. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 551–553, New York, NY, USA, 2019. Association for Computing Machinery.

[28] Darren Guinness, Edward Cutrell, and Meredith Ringel Morris. Caption crawler: Enabling reusable alternative text descriptions using reverse image search. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2018.

[29] Alexandra Vtyurina, Adam Fourney, Meredith Ringel Morris, Leah Findlater, and Ryen W. White. Verse: Bridging screen readers and voice assistants for enhanced eyes-free web search. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 414–426, New York, NY, USA, 2019. Association for Computing Machinery.

[30] Daisuke Sato, Shaojian Zhu, Masatomo Kobayashi, Hironobu Takagi, and Chieko Asakawa. Sasayaki: Augmented voice web browsing experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2769–2778, New York, NY, USA, 2011. ACM.

[31] Shaojian Zhu, Daisuke Sato, Hironobu Takagi, and Chieko Asakawa. Sasayaki: An augmented voice-based web browsing experience. In *Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '10, pages 279–280, New York, NY, USA, 2010. ACM.

[32] Meredith Ringel Morris, Jazette Johnson, Cynthia L. Bennett, and Edward Cutrell. Rich representations of visual content for screen reader users. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–11, New York, NY, USA, 2018. Association for Computing Machinery.

[33] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3195–3204, 2019.

[34] Julian McAuley and Alex Yang. Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of the 25th International Conference on World Wide Web*, pages 625–635, 2016.

[35] Mengting Wan and Julian McAuley. Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 489–498. IEEE, 2016.

[36] Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang, and Hao Yu. Structure-aware review mining and summarization. In *Proceedings of the 23rd international conference on computational linguistics*, pages 653–661. Association for Computational Linguistics, 2010.

[37] Jeff Huang, Oren Etzioni, Luke Zettlemoyer, Kevin Clark, and Christian Lee. Revminer: An extractive interface for navigating reviews on a smartphone. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*, UIST '12, pages 3–12, New York, NY, USA, 2012. ACM.

[38] Hye-Jin Min and Jong C Park. Identifying helpful reviews based on customer's mentions about experiences. *Expert Systems with Applications*, 39(15):11830–11838, 2012.

[39] Quang Nguyen. Detecting experience revealing sentences in product reviews. *University of Amsterdam*, 2012.

[40] Ido Guy, Avihai Mejer, Alexander Nus, and Fiana Raiber. Extracting and ranking travel tips from user-generated reviews. In *Proceedings of the 26th international conference on world wide web*, pages 987–996, 2017.

[41] Di Zhu, Theodoros Lappas, and Juheng Zhang. Unsupervised tip-mining from customer reviews. *Decision Support Systems*, 107:116–124, 2018.

[42] Slava Novgorodov, Ido Guy, Guy Elad, and Kira Radinsky. Generating product descriptions from user reviews. In *The World Wide Web Conference*, WWW '19, page 1354–1364, New York, NY, USA, 2019. Association for Computing Machinery.

[43] Guy Elad, Ido Guy, Slava Novgorodov, Benny Kimelfeld, and Kira Radinsky. Learning to generate personalized product descriptions. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 389–398, New York, NY, USA, 2019. Association for Computing Machinery.

[44] Joachim Wagner, Piyush Arora, Santiago Cortes, Utsab Barman, Dasha Bogdanova, Jennifer Foster, and Lamia Tounsi. Dcu: Aspect-based polarity classification for semeval task 4. 2014.

[45] Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 437–442, 2014.

[46] Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 380–385, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[47] Virginia Braun and Victoria Clarke. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4):589–597, 2019.

[48] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.

[49] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6720–6731, 2019.

[50] Vinicius Woloszyn, Henrique D. P. dos Santos, Leandro Krug Wives, and Karin Becker. Mrr: An unsupervised algorithm to rank reviews by relevance. In *Proceedings of the International Conference on Web Intelligence*, WI '17, page 877–883, New York, NY, USA, 2017. Association for Computing Machinery.

[51] Shari Trewin, Diogo Marques, and Tiago Guerreiro. Usage of subjective scales in accessibility research. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, ASSETS '15, page 59–67, New York, NY, USA, 2015. Association for Computing Machinery.

[52] Emeline Brulé, Brianna J. Tomlinson, Oussama Metatla, Christophe Jouffrais, and Marcos Serrano. Review of quantitative empirical evaluations of technology for people with visual impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery.

[53] Xingyu Liu, Patrick Carrington, Xiang'Anthony' Chen, and Amy Pavel. What makes videos accessible to blind and visually impaired people? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, New York, NY, USA, 2021. Association for Computing Machinery.

[54] Chris Mikul. Audio description background paper. *Media Access Australia*, 2010.

[55] Yujia Wang, Wei Liang, Haikun Huang, Yongqi Zhang, Dingzeyu Li, and Lap-Fai Yu. Toward automatic audio description generation for accessible videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–12, New York, NY, USA, 2021. Association for Computing Machinery.

[56] Xingyu Liu, Ruolin Wang, Dingzeyu Li, Xiang'Anthony' Chen, Amy Pavel, et al. Crossa11y: Identifying video accessibility issues via cross-modal grounding. *arXiv preprint arXiv:2208.11144*, 2022.

[57] Beste F Yuksel, Pooyan Fazli, Umang Mathur, Vaishali Bisht, Soo Jung Kim, Joshua Junhee Lee, Seung Jung Jin, Yue-Ting Siu, Joshua A Miele, and Ilmi Yoon. Human-in-the-loop machine learning to increase video accessibility for visually impaired and blind users. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, pages 47–60, New York, NY, USA, 2020. Association for Computing Machinery.

[58] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. Soylent: A word processor with a crowd inside. *Commun. ACM*, 58(8):85–94, jul 2015.

[59] The Smith-Kettlewell Eye Research Institute. Youdescribe, 2022.

[60] Mina Huh, YunJung Lee, Dasom Choi, Haesoo Kim, Uran Oh, and Juho Kim. Cocomix: Utilizing comments to improve non-visual webtoon accessibility. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[61] Matin Yarmand, Dongwook Yoon, Samuel Dodson, Ido Roll, and Sidney S. Fels. "can you believe [1:21]?!": Content and time-based reference patterns in video comments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery.

[62] Brian J Fogg. A behavior model for persuasive design. In *Proceedings of the 4th international Conference on Persuasive Technology*, pages 1–7, 2009.

[63] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. Vizwiz: Nearly real-time answers to visual questions. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, page 333–342, New York, NY, USA, 2010. Association for Computing Machinery.

[64] Mingrui Ray Zhang, Mingyuan Zhong, and Jacob O. Wobbrock. Ga11y: An automated gif annotation system for visually impaired users. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[65] Sarah Bawabe, Laura Wilson, Tongyu Zhou, Ezra Marks, and Jeff Huang. The ux factor: Using comparative peer review to evaluate designs through user preferences. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021.

[66] Danielle Bragg, Naomi Caselli, John W. Gallagher, Miriam Goldberg, Courtney J. Oka, and William Thies. Asl sea battle: Gamifying sign language data collection. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[67] Guanhong Liu, Xianghua Ding, Chun Yu, Lan Gao, Xingyu Chi, and Yuanchun Shi. "i bought this for me to look more ordinary": A study of blind people doing online shopping. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–11, New York, NY, USA, 2019. Association for Computing Machinery.

[68] Abigale J. Stangl, Esha Kothari, Suyog D. Jain, Tom Yeh, Kristen Grauman, and Danna Gurari. Browsewithme: An online clothes shopping assistant for people with visual impairments. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '18, page 107–118, New York, NY, USA, 2018. Association for Computing Machinery.

[69] Violeta Voykinska, Shiri Azenkot, Shaomei Wu, and Gilly Leshed. How blind people interact with visual content on social networking services. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work amp; Social Computing*, CSCW '16, page 1584–1595, New York, NY, USA, 2016. Association for Computing Machinery.

[70] Langis Gagnon, Samuel Foucher, Maguelonne Heritier, Marc Lalonde, David Byrns, Claude Chapdelaine, James Turner, Suzanne Mathieu, Denis Laurendeau, Nath Tan Nguyen, et al. Towards computer-vision software tools to increase production and accessibility of video description for people with vision loss. *Universal Access in the Information Society*, 8(3):199–218, 2009.

[71] Beste F. Yuksel, Soo Jung Kim, Seung Jung Jin, Joshua Junhee Lee, Pooyan Fazli, Umang Mathur, Vaishali Bisht, Ilmi Yoon, Yue-Ting Siu, and Joshua A. Miele. Increasing video accessibility for visually impaired users with human-in-the-loop machine learning. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–9, New York, NY, USA, 2020. Association for Computing Machinery.

[72] Beste F. Yuksel, Pooyan Fazli, Umang Mathur, Vaishali Bisht, Soo Jung Kim, Joshua Junhee Lee, Seung Jung Jin, Yue-Ting Siu, Joshua A. Miele, and Ilmi Yoon. Human-in-the-loop machine learning to increase video accessibility for visually impaired and blind users. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, DIS '20, page 47–60, New York, NY, USA, 2020. Association for Computing Machinery.

[73] Yue Chen, Qin Gao, and Ge Gao. Timeline-anchored comments in video-based learning: The impact of visual layout and content depth. *International Journal of Human–Computer Interaction*, 38(9):868–883, 2022.

[74] Ching-Ying Sung, Xun-Yi Huang, Yicong Shen, Fu-Yin Cherng, Wen-Chieh Lin, and Hao-Chuan Wang. Exploring online learners' interactive dynamics by visually analyz-

ing their time-anchored comments. In *Computer Graphics Forum*, volume 36, pages 145–155. Wiley Online Library, 2017.

[75] Ching-Ying Sung, Hao-Chuan Wang, Chen-Wei Huang, and Wen-Chieh Lin. Questionization of peer-generated comments for supporting online video-based learning. In *Proceedings of the Fifth International Symposium of Chinese CHI*, pages 11–15, 2017.

[76] Xiaojuan Ma and Nan Cao. Video-based evanescent, anonymous, asynchronous social interaction: Motivation and adaption to medium. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 770–782, New York, NY, USA, 2017. Association for Computing Machinery.

[77] Woosuk Seo and Hyunggu Jung. Understanding the community of blind or visually impaired vloggers on youtube. *Universal Access in the Information Society*, 20:31–44, 2021.

[78] American Council of the Blind. American council of the blind, audio description project, guidelines for audio describers. `https://www.acb.org/adp/guidelines.html`, 2020.

[79] Jaclyn Packer, Katie Vizenor, and Joshua A Miele. An overview of video description: history, benefits, and guidelines. *Journal of Visual Impairment & Blindness*, 109(2):83–93, 2015.

[80] Ben Caldwell, Michael Cooper, Loretta Guarino Reid, Gregg Vanderheiden, Wendy Chisholm, John Slatin, and Jason White. Web content accessibility guidelines (wcag) 2.0. *WWW Consortium (W3C)*, 290:1–34, 2008.

[81] Terril Thompson. My audio description talk @ csun, 2017.

[82] Masatomo Kobayashi, Trisha O'Connell, Bryan Gould, Hironobu Takagi, and Chieko Asakawa. Are synthesized video descriptions acceptable? In *Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '10, page 163–170, New York, NY, USA, 2010. Association for Computing Machinery.

[83] Rosiana Natalie, Jolene Loh, Huei Suen Tan, Joshua Tseng, Ian Luke Yi-Ren Chan, Ebrima H Jarjue, Hernisa Kacorri, and Kotaro Hara. The efficacy of collaborative authoring of video scene descriptions. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '21, New York, NY, USA, 2021. Association for Computing Machinery.

[84] Carmen J Branje and Deborah I Fels. Livedescribe: can amateur describers create high-quality audio description? *Journal of Visual Impairment & Blindness*, 106(3):154–165, 2012.

[85] L. Gagnon, C. Chapdelaine, D. Byrns, S. Foucher, M. Héritier, and V. Gupta. A computer-vision-assisted system for videodescription scripting. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 41–48, New York, NY, USA, 2010. Institute of Electrical and Electronics Engineers.

[86] Virginia P Campos, Tiago MU de Araújo, Guido L de Souza Filho, and Luiz MG Gonçalves. Cinead: a system for automated audio description script generation for the visually impaired. *Universal Access in the Information Society*, 19(1):99–111, 2020.

[87] 3PlayMedia. 3playmedia, 2022.

[88] Masatomo Kobayashi, Kentarou Fukuda, Hironobu Takagi, and Chieko Asakawa. Providing synthesized audio description for online videos. In *Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets '09, page 249–250, New York, NY, USA, 2009. Association for Computing Machinery.

[89] Jeffrey P Bigham and Kyle Murray. Webtrax: visualizing non-visual web interactions. In *Computers Helping People with Special Needs: 12th International Conference, ICCHP 2010, Vienna, Austria, July14-16, 2010, Proceedings, Part II 12*, pages 346–353. Springer, 2010.

[90] Daisuke Sato, Masatomo Kobayashi, Hironobu Takagi, and Chieko Asakawa. What's next? a visual editor for correcting reading order. In *Human-Computer Interaction–INTERACT 2009: 12th IFIP TC 13 International Conference, Uppsala, Sweden, August 24-28, 2009, Proceedings, Part I 12*, pages 364–377. Springer, 2009.

[91] Hironobu Takagi, Chieko Asakawa, Kentarou Fukuda, and Junji Maeda. Accessibility designer: visualizing usability for the blind. *ACM SIGACCESS accessibility and computing*, (77-78):177–184, 2003.

[92] Yi-Hao Peng, JiWoong Jang, Jeffrey P Bigham, and Amy Pavel. Say it all: Feedback for improving non-visual presentation accessibility. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[93] Valerie S. Morash, Yue-Ting Siu, Joshua A. Miele, Lucia Hasty, and Steven Landau. Guiding novice web workers in making image descriptions using templates. *ACM Trans. Access. Comput.*, 7(4), nov 2015.

[94] Masaki Kobayashi, Hiromi Morita, Masaki Matsubara, N. Shimizu, and Atsuyuki Morishima. An empirical study on short- and long-term effects of self-correction in crowdsourced microtasks. In *HCOMP*, 2018.

[95] Rachel N. Simons, Danna Gurari, and Kenneth R. Fleischmann. "i hope this is helpful": Understanding crowdworkers' challenges and motivations for an image description task. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), oct 2020.

[96] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Comput. Surv.*, 51(1), jan 2018.

[97] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–15, New York, NY, USA, 2019. Association for Computing Machinery.

[98] Farnaz Jahanbakhsh, Amy X. Zhang, Adam J. Berinsky, Gordon Pennycook, David G. Rand, and David R. Karger. Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), April 2021.

[99] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. The future of crowd work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, page 1301–1318, New York, NY, USA, 2013. Association for Computing Machinery.

[100] Jaime Teevan. The future of microwork. *XRDS*, 23(2):26–29, dec 2016.

[101] Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. Cascade: crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 1999–2008, New York, NY, USA, 2013. Association for Computing Machinery.

[102] Jaime Teevan, Shamsi T. Iqbal, and Curtis von Veh. Supporting collaborative writing with microtasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 2657–2668, New York, NY, USA, 2016. Association for Computing Machinery.

[103] Jaime Teevan, Daniel J. Liebling, and Walter S. Lasecki. Selfsourcing personal tasks. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14, page 2527–2532, New York, NY, USA, 2014. Association for Computing Machinery.

[104] Amy X. Zhang and Justin Cranshaw. Making sense of group chat through collaborative tagging and summarization. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):196:1–196:27, November 2018.

[105] Walter S. Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F. Allen, and Jeffrey P. Bigham. Chorus: a crowd-powered conversational assistant. In *Proceedings*

of the 26th Annual ACM Symposium on User Interface Software and Technology, UIST '13, page 151–162, New York, NY, USA, 2013. Association for Computing Machinery.

[106] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. Calendar. help: Designing a workflow-based scheduling agent with humans in the loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2382–2393, 2017.

[107] Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. Real-time captioning by groups of non-experts. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*, UIST '12, page 23–34, New York, NY, USA, 2012. Association for Computing Machinery.

[108] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, New York, NY, USA, 2020. Institute of Electrical and Electronics Engineers.

[109] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33:25–37, 2020.

[110] Graham R Gibbs. Thematic coding and categorizing. *Analyzing qualitative data*, 703:38–56, 2007.

[111] Soyoung Park. A study on visual scaffolding design principles in web-based learning environments. *Electronic Journal of E-Learning*, 20(2):pp180–200, 2022.

[112] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. From game design elements to gamefulness: Defining "gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, MindTrek '11, page 9–15, New York, NY, USA, 2011. Association for Computing Machinery.

[113] Kelly Mack, Edward Cutrell, Bongshin Lee, and Meredith Ringel Morris. Designing tools for high-quality alt text authoring. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '21, New York, NY, USA, 2021. Association for Computing Machinery.

[114] Stacy M. Branham and Shaun K. Kane. Collaborative accessibility: How blind and sighted companions co-create accessible home spaces. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 2373–2382, New York, NY, USA, 2015. Association for Computing Machinery.

[115] Stacy M. Branham and Shaun K. Kane. The invisible work of accessibility: How blind employees manage accessibility in mixed-ability workplaces. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, ASSETS '15, page 163–171, New York, NY, USA, 2015. Association for Computing Machinery.

[116] Ethan Z. Rong, Mo Morgana Zhou, Zhicong Lu, and Mingming Fan. "it feels like being locked in a cage": Understanding blind or low vision streamers' perceptions of content curation algorithms. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, DIS '22, page 571–585, New York, NY, USA, 2022. Association for Computing Machinery.

[117] Franklin Mingzhe Li, Di Laura Chen, Mingming Fan, and Khai N. Truong. "i choose assistive devices that save my face": A study on perceptions of accessibility and assistive technology use conducted in china. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[118] Cass R Sunstein. Nudges that fail. *Behavioural public policy*, 1(1):4–25, 2017.

[119] Lucy Lu Wang, Isabel Cachola, Jonathan Bragg, Evie Yu-Yen Cheng, Chelsea Haupt, Matt Latzke, Bailey Kuehl, Madeleine van Zuylen, Linda Wagner, and Daniel S Weld. Improving the accessibility of scientific documents: Current state, user needs, and a system solution to enhance scientific pdf accessibility for blind and low vision users. *arXiv preprint arXiv:2105.00076*, 2021.

[120] Sanjana Shivani Chintalapati, Jonathan Bragg, and Lucy Lu Wang. A dataset of alt texts from hci publications: Analyses and uses towards producing more descriptive alt texts of data visualizations in scientific papers. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '22, New York, NY, USA, 2022. Association for Computing Machinery.

[121] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.

[122] Xin Qian, Eunyee Koh, Fan Du, Sungchul Kim, and Joel Chan. A formative study on designing accurate and natural figure captioning systems. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–8, New York, NY, USA, 2020. Association for Computing Machinery.

[123] Elliot Salisbury, Ece Kamar, and Meredith Ringel Morris. 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*, 2017.

[124] Steven Komarov and Krzysztof Z Gajos. Organic peer assessment. In *Proceedings of the CHI 2014 Learning Innovation at Scale workshop*, 2014.

[125] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J. Guo, Robert C. Miller, and Krzysztof Z. Gajos. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 4017–4026, New York, NY, USA, 2014. Association for Computing Machinery.

[126] Pao Siangliulue, Joel Chan, Steven P. Dow, and Krzysztof Z. Gajos. Ideahound: Improving large-scale collaborative ideation with crowd-powered real-time semantic modeling. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, page 609–624, New York, NY, USA, 2016. Association for Computing Machinery.

[127] Gina Biancarosa and Gina G Griffiths. Technology tools to support reading in the digital age. *The future of children*, pages 139–160, 2012.

[128] Abigail J Sellen and Richard HR Harper. *The myth of the paperless office*. MIT press, 2003.

[129] Kenton O'Hara and Abigail Sellen. A comparison of reading paper and on-line documents. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, CHI '97, page 335–342, New York, NY, USA, 1997. Association for Computing Machinery.

[130] Sheng-Shiang Tseng, Hui-Chin Yeh, and Shih-hsien Yang. Promoting different reading comprehension levels through online annotations. *Computer Assisted Language Learning*, 28(1):41–57, 2015.

[131] Hui-Chin Yeh, Hsiu-Ting Hung, and Yu-Hsin Chiang. The use of online annotations in reading instruction and its impact on students' reading progress and processes. *ReCALL*, 29(1):22–38, 2017.

[132] Francis P Robinson. Effective study, rev. 1946.

[133] Thomas H Anderson and Bonnie B Armbruster. Reader and text-studying strategies. 1982.

[134] Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs' Ka, Xiang'Anthony' Chen, and Caiming Xiong. Designing and evaluating interfaces that highlight news coverage diversity using discord questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2023.

[135] An T Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C Wallace, and Matthew Lease. Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking. In *Proceedings of the 31st annual ACM symposium on user interface software and technology*, pages 189–199, 2018.

[136] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. Newscube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 443–452, 2009.

[137] Eric Balagtas Perez, James King, Yugo H Watanabe, and Xiang'Anthony' Chen. Counterweight: Diversifying news consumption. In *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 132–134, 2020.

[138] Xiang "Anthony" Chen, Chien-Sheng Wu, Lidiya Murakhovs' ka, Philippe Laban, Tong Niu, Wenhao Liu, and Caiming Xiong. Marvista: exploring the design of a human-ai collaborative news reading tool. *ACM Transactions on Computer-Human Interaction*, 30(6):1–27, 2023.

[139] Srinivasan Keshav. How to read a paper. *ACM SIGCOMM Computer Communication Review*, 37(3):83–84, 2007.

[140] Terry Tomasek. Critical reading: Using reading prompts to promote active engagement with text. *International journal of teaching and learning in higher education*, 21(1):127–132, 2009.

[141] Philip C Abrami, Robert M Bernard, Evgueni Borokhovski, Anne Wade, Michael A Surkes, Rana Tamim, and Dai Zhang. Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of educational research*, 78(4):1102–1134, 2008.

[142] Roland Case. Moving critical thinking to the main stage. *Education Canada*, 45(2):45–49, 2005.

[143] Richard Paul and Linda Elder. Critical thinking: What, why, and how. *New directions for community colleges*, 77(2):3–24, 1992.

[144] Diane F Halpern. Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American psychologist*, 53(4):449, 1998.

[145] Kangyu Yuan, Hehai Lin, Shilei Cao, Zhenhui Peng, Qingyu Guo, and Xiaojuan Ma. Critrainer: An adaptive training tool for critical paper reading. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery.

[146] Christin Jonathan, Jennifer Pei-Ling Tan, Elizabeth Koh, Imelda S Caleon, Siu Hua Tay, W Chen, JC Yang, AF Ayub, SL Wong, and A Mitrovic. Enhancing students' critical reading fluency, engagement and self-efficacy using self-referenced learning analytics dashboard visualizations. In *Proceedings of the 25th international conference on computers in education*, pages 457–462. Asia-Pacific Society for Computers in Education Jhongli, Taiwan, 2017.

[147] Zhenhui Peng, Yuzhi Liu, Hanqi Zhou, Zuyu Xu, and Xiaojuan Ma. Crebot: Exploring interactive question prompts for critical paper reading. *International Journal of Human-Computer Studies*, 167:102898, 2022.

[148] Adam Strantz. Using web standards to design accessible data visualizations in professional communication. *IEEE Transactions on Professional Communication*, 64(3):288–301, 2021.

[149] Michael D Crossland, Rui S. Silva, and Antonio F Macedo. Smartphone, tablet computer and e-reader use by people with vision impairment. *Ophthalmic and Physiological Optics*, 34(5):552–557, 2014.

[150] Edward C Summers, Julianna Elizabeth Langston, Robert Allison, and Jennifer Cowley. Using sas/graph ® to create visualizations that also support tactile and auditory interaction. 2012.

[151] Meredith Ringel Morris, Jazette Johnson, Cynthia L. Bennett, and Edward Cutrell. Rich representations of visual content for screen reader users. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[152] Margot Hanley, Solon Barocas, Karen Levy, Shiri Azenkot, and Helen Nissenbaum. Computer vision and conflicting values: Describing people with automated alt text. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 543–554, 2021.

[153] Soya Park, Jonathan Bragg, Michael Chang, Kevin Larson, and Danielle Bragg. Exploring team-sourced hyperlinks to address navigation challenges for low-vision readers of scientific papers. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), nov 2022.

[154] Zhuohao Zhang, Gene S-H Kim, and Jacob O. Wobbrock. Developing and deploying a real-world solution for accessible slide reading and authoring for blind users. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '23, New York, NY, USA, 2023. Association for Computing Machinery.

[155] Zhuohao Zhang and Jacob O. Wobbrock. A11yboard: Using multimodal input and output to make digital artboards accessible to blind users. In *Adjunct Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22 Adjunct, New York, NY, USA, 2022. Association for Computing Machinery.

[156] Zhuohao (Jerry) Zhang and Jacob O. Wobbrock. A11yboard: Making digital artboards accessible to blind and low-vision users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[157] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S. Weld, and Marti A. Hearst. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[158] Alan Lundgard and Arvind Satyanarayan. Accessible Visualization via Natural Language Descriptions: A Four-Level Model of Semantic Content. *IEEE Transactions on Visualization & Computer Graphics (Proc. IEEE VIS)*, 2022.

[159] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.

[160] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

[161] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):1–26, 2024.

[162] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

[163] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

[164] Yiran Wu, Feiran Jia, Shaokun Zhang, Qingyun Wu, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, and Chi Wang. An empirical study on challenging math problem solving with gpt-4. *arXiv preprint arXiv:2306.01337*, 2023.

[165] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023.

[166] Zongjie Li, Chaozheng Wang, Chaowei Liu, Pingchuan Ma, Daoyuan Wu, Shuai Wang, and Cuiyun Gao. Vrptest: Evaluating visual referring prompting in large multimodal models. *arXiv preprint arXiv:2312.04087*, 2023.

[167] Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*, 2024.

[168] Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*, 2023.

[169] Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.

[170] Muhammad Aurangzeb Ahmad, Ilker Yaramis, and Taposh Dutta Roy. Creating trustworthy llms: Dealing with hallucinations in healthcare ai. *arXiv preprint arXiv:2311.01463*, 2023.

[171] Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. Generation-augmented retrieval for open-domain question answering. *arXiv preprint arXiv:2009.08553*, 2020.

[172] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762, 2024.

[173] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[174] Maitraye Das, Anne Marie Piper, and Darren Gergle. Design and evaluation of accessible collaborative writing techniques for people with vision impairments. *ACM Trans. Comput.-Hum. Interact.*, 29(2), jan 2022.

[175] Maitraye Das, Thomas Barlow McHugh, Anne Marie Piper, and Darren Gergle. Co11ab: Augmenting accessibility in synchronous collaborative writing for people with vision impairments. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[176] Venkatesh Potluri, Maulishree Pandey, Andrew Begel, Michael Barnett, and Scott Reitherman. Codewalk: Facilitating shared awareness in mixed-ability collaborative software development. In *Proceedings of the 24th International ACM SIGACCESS*

Conference on Computers and Accessibility, ASSETS '22, New York, NY, USA, 2022. Association for Computing Machinery.

[177] Ali Abdolrahmani, Maya Howes Gupta, Mei-Lian Vader, Ravi Kuber, and Stacy Branham. Towards more transactional voice assistants: Investigating the potential for a multimodal voice-activated indoor navigation assistant for blind and sighted travelers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[178] Michael S Bernstein. Crowd-powered systems. *KI-Künstliche Intelligenz*, 27:69–73, 2013.

[179] Kat Holmes. *Mismatch: How inclusion shapes design*. Mit Press, 2020.

[180] Aoyu Wu, Wai Tong, Haotian Li, Dominik Moritz, Yong Wang, and Huamin Qu. Computableviz: Mathematical operators as a formalism for visualisation processing and analysis. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.